

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



SCIENCE SHARED

When communities and
researchers work together

PAGE 23

ASTROPHYSICS

THE HUNT FOR DARK MATTER

How to find the most elusive
particles in the Universe

PAGE 51

MECHANOBIOLOGY

LIVERS FEEL THE FORCE

Blood flow helps to drive
hepatocyte proliferation

PAGES 42 & 128

ARCHAEOLOGY

PREHISTORIC ABSTRACT ART

73,000-year-old pattern is
earliest known drawing

PAGE 115

NATURE.COM

4 October 2018

Vol. 562, No. 7725

THIS WEEK

EDITORIALS

TASTE Savour the flavour of a gene-edited tomato **p.8**

WORLD VIEW How farmers transformed climate-science project **p.9**



YELLOWSTONE More endangered than your average bear **p.13**

Power to the people

Everyone gains when researchers partner with the public and policymakers. The knowledge generated is more likely to be useful to society and should be encouraged.

Few sign up to science for a glamorous lifestyle, colossal salary or generous dental plan. They do it to gain insights and knowledge and, they hope, to make the world a better place. Too often, that last objective proves hard to achieve — not because of uncaring researchers living in ivory towers, but because the way in which some types of study are done and rewarded does not set the correct priorities. That needs to change.

Enter co-production: full involvement in research by people who hope to benefit from the work, in partnership with communities, policymakers and other members of the public. Popular since the 1970s among sociologists as a way to help set inclusive policy, the term — and the principle — is spreading throughout academic science. As we highlight in a special issue this week, a growing number of projects are adopting the approach and working with such groups to jointly carry out research. And ‘jointly’ applies at all stages, from the project’s initial framing through to publication and follow-up.

Co-production can take many forms. Climate scientists, for example, are partnering with farmers to tailor projects to focus on their specific circumstances, such as the changes in precipitation that are likely in a warming world. A World View on page 9 explores how researchers worked with farmers in northeast Argentina to produce forecast systems for local needs, on the basis of emerging climate models and local knowledge of crop losses.

Clinicians, environmental researchers and many others are coming to appreciate that there are crucial kinds of lived expertise that can improve their studies.

Co-production is less suited to some scientific pursuits, but it can be a powerful way to make results more relevant and practicable across a spread of disciplines. Some call it science that is actionable. At present, too much research done in the name of society is not used by society. Instead it is paid for, produced and dutifully recorded, and then left waiting for someone to come along and use it.

Co-production demands a different approach — from funders, who need to find flexible ways to include and pay for people who work outside academia, to institutions, some of which appoint dedicated staff to negotiate and champion the sometimes-sensitive partnerships required. It needs better incentives: ones that recognize that this work often takes time and doesn’t necessarily lead to high-profile papers and other conventional types of academic success, but can produce outcomes that make a difference in the lives of the people at the heart of the research. Also needed are better ways to analyse and measure the success of co-produced research (see Comment, page 32). Publishers and journals can play a part; in one small

step, for instance, *Nature*’s authorship guidelines state that anyone who had a sufficient role in the

work can be included as an author (see go.nature.com/2pocpux). Most of all, co-production requires individual scientists to see the opportunities and to want to take advantage of them.

The growth in political populism and rising public dissatisfaction with policies some people see as excluding their interests have made it more important for researchers to produce — and to be seen to produce — research that is both beneficial and relevant to society. Efforts to do so are overdue. The onus is on researchers and those who support them to put systems in place to encourage more collaborations.

If ivory-tower scientists did cut themselves off from the problems of the world in the past (and multiple lines of evidence over decades across medicine, engineering, technology, agriculture and dozens of other fields suggest that many did not), then few can get away with such an attitude now. Grant applications and project assessments ask for explanations of the work’s probable societal impact, and commercial funding frequently comes with a desired application as a goal.

Co-production is better for society. It also leads to better research — both technically, because it accounts for more factors, and ethically, because it’s more equitable. That means it increases the chances of genuinely making the world a better place, because what emerges will be more suitable for take-up. That’s something that everyone who cares about research can sign up to. ■

Forgotten crime

The United States should not execute a murderer who no longer remembers his offence.

In 1985, Vernon Madison shot a police officer in the back of the head while the officer sat in his car. It was a heinous crime for which Madison has spent the past 33 years on death row, much of it in solitary confinement.

But Madison now doesn’t remember the Alabama shooting or the name of the officer — he can remember very little at all. Multiple strokes in the past few years have wiped out parts of his brain involved in memory and left him with vascular dementia. Madison’s lawyers have appealed against his death sentence, and presented his case before the US Supreme Court earlier this week.

The case raises complex philosophical, legal and ethical questions about the purpose of the death penalty and what it means to truly understand one’s own guilt. In taking the case, the court accepted the task of deciding whether it is cruel and unusual to execute a violent murderer



CO-PRODUCTION OF RESEARCH
A *Nature* special issue
nature.com/collections/coproduction

who doesn't understand his fate, even if he understood the possible consequences at the time the crime occurred.

The evidence of Madison's cognitive disability is convincing. In neuropsychological tests administered by multiple specialists, he can't interpret the meaning of stories or logically draw conclusions. His lawyers say that, in terms of his intellectual function, there is no difference between his current condition and that of a person born with an intellectual disability. The latter group is protected from execution, thanks to a 2002 Supreme Court decision.

Madison's case differs because he did not have a severe cognitive impairment at the time he committed the murder, and presumably knew it was wrong. The state of Alabama argues that once the situation is explained to him, Madison also understands that he was tried and will be executed. Alabama says it doesn't matter whether he remembers it, because he can still rationally conceptualize it.

But psychologists and psychiatrists say that this is very different from a deep understanding of one's own guilt. Ultimately, the court will have to determine what level of 'understanding' is sufficient to conclude that Madison can rationally process his punishment.

Science cannot offer all the answers in this specific case. Still, decades of research on neuropsychology — much of it done to better understand mental-health conditions — have honed the ability of specialists to understand brain function. In a joint amicus brief filed to the Supreme

Court, the American Psychiatric Association and the American Psychological Association say that neuropsychological tests and advances in neuroimaging can accurately assess cognitive capacity, precluding any concerns that in the future, courts might see a flood of appeals from people who falsely claim to have no memory of their crime.

The court could issue a narrow decision that applies only to Madison, or it could rule broadly on people who cannot understand the reason for punishment. Although there are probably very few people on death row with vascular dementia, conditions such as traumatic brain injury or tumours could cause a person to forget a crime.

“The case highlights the illogic of capital punishment.”

The case highlights the illogic of capital punishment. Death-penalty proponents argue that it is necessary for justice to be served, as well as to deter others from crime. Yet neither of these conditions applies here. Madison cannot see his execution as justice because he cannot recall his crime. And executing a person with an intellectual disability hardly serves as an example or deterrent.

Regardless of the decision, Madison is not going unpunished. If he escapes execution, he will spend the rest of his life in prison alone, disabled and confused by the world around him. He is no longer a threat. The court should set an example and grant mercy. ■

I say tomato

A new super-tomato highlights Europe's outdated approach to gene editing.

The world produces some 800 billion tomatoes each year — but how many of them are worth eating? Thousands of years of breeding have produced a fruit that often suits farmers and sellers more than consumers. Vines now grow in an orderly fashion, and produce lots of tomatoes that stay in place until they are harvested and are firm enough to be shipped long distances. But, in too many cases, studies have confirmed that flavour and nutrition have got lost somewhere along the way.

Plant scientists are on the case. This week, three papers from research groups around the world detail attempts to make a new type of super-tomato: one that does not sacrifice taste for convenience. To do this, the researchers used CRISPR–Cas9 gene editing, which allowed them to modify specific genes in wild relatives of tomatoes. The result — according to a scientist who has tasted one of the fruits — is an “aromatic” tomato that could re-energize taste buds.

The studies are a demonstration of the fruits of decades of painstaking plant-genetics research: a cupboard full of genes with known effects, that can each be adjusted to turn an unruly wild plant into a valuable domesticated one. The work serves as a reminder of the value of basic research into plant growth and development. And it shows how other useful traits could be introduced in other crops.

One group edited a wild relative of the tomato called *Physalis peruviana*, which is grown in Central and South America (Z. H. Lemmon *et al. Nature Plants* 4, 766–770; 2018). Its berries are tasty and slightly sweet, but its sprawling growth pattern and tendency to drop its fruit onto the ground make it ill-suited for large-scale agriculture. The edited plant was more compact, and produced larger fruits.

The other two groups tinkered with a relative called *Solanum pimpinellifolium*. This species is stress tolerant and resistant to the commercially devastating disease bacterial spot, but the researchers sought to boost the size and attractiveness of its fruits, while making plant growth easier to control (A. Zsögön *et al. Nature Biotechnol.* <http://doi.org/cvf2>; 2018; T. Li *et al. Nature*

Biotechnol. <http://doi.org/cvfz>; 2018). They aimed to combine the benefits of *S. pimpinellifolium* with the features of modern tomatoes that appeal to farmers and consumers. The researchers also laboured to increase the nutritional value of their new tomatoes: first, by boosting the levels of lycopene, a carotenoid linked to health benefits; and second, by focusing on a greater vitamin C content.

To achieve the same product through conventional breeding would have taken decades, says Jörg Kudla at the University of Münster in Germany, a lead author on one of the papers. Instead, it took his team three years. It's an example of science serving a need of society — and one that highlights the flawed steps the European Union is taking that will threaten such work in the future. In July, the European Court of Justice ruled that foods produced by CRISPR–Cas9 gene editing must be bound by the same onerous regulations as genetically modified crops. The resulting mandatory tests and trials will massively increase the cost of developing a commercial product, which in turn makes funding for research on such products less viable.

The expense is one reason why genetically modified crops have so far yielded little benefit for consumers: because it has cost so much to produce such plants, companies focus on developing commodity crops and traits that appeal to farmers. Kudla has grant applications for up to €2 million (US\$2.3 million) now under review to fund research related to his gene-editing work. But funders have a responsibility to spend their cash in ways that might benefit taxpayers, he notes, and if such crops have no commercial future in Europe, it might be a struggle to justify paying for the crops' development.

The long-awaited European court decision puzzled many researchers, because the technique involves gene edits that merely disable a gene, rather than rewriting it with a specific sequence. Scientifically, advocates see this as being similar to using a chemical or radiation to generate mutations and then screening the plants for a desired trait — which is not classed as genetic modification. But with CRISPR–Cas9, researchers can generate the mutations in specific genes, without having to screen thousands of plants for each trait they want to introduce.

The ruling came as a blow, particularly because, in January, an advocate-general to the European court argued that such crops do not need the same scrutiny as conventional genetically modified crops. And it highlights the degree to which researchers are at odds with officials on genetic modification in Europe. Scientists and supporters must keep up their efforts to advocate for cutting-edge research. Meanwhile, perhaps a better-tasting tomato could help to bring more policymakers on side. ■



Farmers transformed how we investigate climate

To make my research more useful for people deciding how to plant crops and prevent flood damage, I asked for their help, says Carolina Vera.

People living in the Matanza River basin in eastern Argentina know where to expect floods. When they pooled their collective experience to make a flood-risk map, the result was essentially the same as what we scientists would make with hydrology and watershed measurements. With scientists and locals working together, we matched up past floods with the rainfall that produced them and pinned down how a small amount of precipitation can have a huge effect. Now we frame our research in that region around the potential for impact, alongside scientifically common measures (such as extreme precipitation).

Too often, we scientists assume that knowledge transfer begins with basic research, which then inspires an application that we expect others to use automatically. Experience has taught me how limited this linear model can be. Co-production — working with those who will actually use the outcomes of my climate research — can be circuitous and unpredictable, but ultimately is more worthwhile.

Over the past two decades, my research and that of my colleagues has helped to build a better picture of the climate of southern South America. Our models now account for the South American monsoon and other mechanisms associated with ocean–atmosphere–land interactions, such as how moisture moves from the tropical Atlantic Ocean to the Amazon basin and farther south.

Initially, we built tools to predict climate weeks or months in advance and set up a website to make that information freely available. We thought it could be valuable for decisions around generating hydroelectric energy, planning for floods, and scheduling plantings and harvests. We did plenty of outreach and encouraged feedback and questions. But few outside academia actually used this information. Why was it so hard to make our climate knowledge more useful?

In 2009, I was selected as a lead author on a report of the Intergovernmental Panel on Climate Change about how communities can manage the risks associated with extreme weather. That gave me the opportunity to work with social scientists, and I realized that making climate science useful also takes social, cultural, economic and even political knowledge. More than that, I needed a dialogue with those who might use or benefit from my research, and to work with them as equals.

In 2016, farmers, anthropologists and climate scientists embarked on a collaboration to develop climate information that would be useful for smallholders who raise cattle and grow vegetables, maize (corn) and potatoes. We had grasped by then that it is not enough to go to the countryside and start to talk. You have to get to know each other and build trust. Our work builds on two years of fieldwork carried out by anthropologists in the Bermejo region of

Chaco province in northeastern Argentina. The doctoral students lived among farmers who must frequently contend with heavy rainfall, hail, frost, strong winds and floods with few resources for recovery.

Our project merged concepts from three domains: climate science (climatic variability, uncertainty, monitoring, normality and predictions); anthropology (perception, confidentiality and participation); and the knowledge that farming communities have earned through experience (direct observation, the occurrence of certain meteorological events in a certain combination and rainfall thresholds that are relevant for production).

We were able to identify phenomena that are not considered extreme by statistical metrics but that have a high impact on the people who live with them. Our collaborators told us about an unusual sequence of more than 15 cloudy days in 2016 that caused winter production of peppers and tomatoes to fall greatly.

We also spotted surprise opportunities. We co-designed a network to monitor rainfall and learn how its spatial distribution determines floods and droughts. Students attending local schools have installed rain gauges on their family farms. They record the data on paper and then upload them at school to an online repository.

I have come to appreciate just how vulnerable communities are. In October 2017, we celebrated the first installations of our rain-observation network. The next month, we lost part of that work when a severe storm blew the roof off the school. Students had to attend classes in a fire station instead. But a year on, we have more rain gauges than before. We are also co-producing a smart-

phone app with local people (see go.nature.com/2noz6k). They use it to see how temperature and precipitation have evolved in the region, along with predictions for the next days and weeks.

I must confess that I have often felt frustrated with the co-production process. The flexibility that enabled the rain-monitoring network and app also makes planning firm schedules and deliverables tough. Dialogue between academic scientists and those who provide weather forecasting or assist agricultural management can be slow, complex and difficult. And it has been hard for me to grasp that others struggle to accept the inherently chaotic nature of the climate, and the impossibility of predicting it precisely.

This experience has changed me as a scientist and as a person. Before, I was a climate researcher strongly motivated to contribute to society, but with hazy notions of how to do so. Now I am part of a process that truly benefits real people as they go about their daily lives. ■

TOO OFTEN, WE
**SCIENTISTS
ASSUME**
THAT KNOWLEDGE
TRANSFER STARTS WITH
**BASIC
RESEARCH.**



CO-PRODUCTION OF RESEARCH
A Nature special issue
nature.com/collections/coproduction

Carolina Vera is a professor at the University of Buenos Aires-CONICET and principal investigator of a co-production project (www.climax-sa.org) to strengthen climate services.
e-mail: carolina@cima.fcen.uba.ar

SEVEN DAYS

The news in brief

POLICY

Misconduct report

A report into how universities in the United Kingdom deal with sexual misconduct reveals “serious causes for concern”. The report, published by lobby organization the 1752 Group, interviewed 16 women who had experienced sexual misconduct from an academic staff member — defined as including grooming, sexual harassment, stalking or sexualized communication — and found that in only one case did the staff member involved lose their job. Many interviewees experienced retaliation, bullying, further trauma, and physical and mental-health problems during what they described as inadequate investigations by institutions into allegations. An analysis of relevant policies from 25 institutions found that some lacked detail on the procedure for handling complaints. The report calls for urgent amendment of common and problematic wording in policy documents that suggests students are responsible for avoiding abuses of power by staff.

SPACE

Observatory sign-up

Ireland is joining the European Southern Observatory (ESO), becoming the 16th member state to join the partnership. Officials from the Irish government and the ESO signed the agreement in Dublin on 26 September. The decision gives Irish astronomers access to the ESO's suite of world-class telescopes in Chile. That includes what will be the largest ground-based telescope in the world, the 39-metre Extremely Large Telescope, which is under construction and expected to come online as early as 2024. Starting in the 1840s, Ireland was home to a 1.8-metre

telescope in Birr, then known as Parsonstown, which was the largest telescope in the world at the time.

Quantum CubeSat

Scientists in the United Kingdom and Singapore plan to launch a quantum satellite in late 2021, they announced on 27 September. The satellite, a lightweight CubeSat, will be designed to demonstrate quantum key distribution (QKD), a technology that harnesses the quantum properties of light to provide inherently secure communication. Together, the two countries' governments plan to invest 18 million Singapore dollars

(US\$13 million) in the project, known as QKD Qubesat, which will transmit secure messages between ground stations in the United Kingdom and Singapore. In 2016, a team led by Pan Jianwei at the University of Science and Technology of China in Hefei became the first to achieve satellite-based QKD, which significantly boosts the distance over which secure communication can happen.

Far-off world

Astronomers have discovered a distant world in the outer Solar System. Named 2015 TG387 and nicknamed The Goblin, it never gets any closer to the Sun than about

65 times the Earth–Sun distance, or roughly twice the current distance between Pluto and the Sun. At its farthest, the object ranges out to 2,300 times the Earth–Sun distance. 2015 TG387 is one of only a handful of objects known in these distant realms. Its orbit is consistent with, but does not prove, the existence of a proposed big planet in the distant Solar System that is popularly known as Planet Nine. A team led by Scott Sheppard, at the Carnegie Institution for Science in Washington DC, discovered The Goblin using Japan's 8.2-metre Subaru telescope on Mauna Kea, Hawaii. The researchers announced the



BAY ISMOYO/AFP/GETTY

Earthquake and tsunami rock Indonesia

A magnitude-7.5 earthquake, and the tsunami it caused, has killed at least 840 people on the Indonesian island of Sulawesi. The quake struck on 28 September about 80 kilometres north of the city of Palu, triggering a tsunami that raced up a narrow inlet and into Palu. No warning sirens

apparently sounded along the beach, where many people were caught in the enormous waves. Buildings collapsed across the region, which also includes the city of Donggala. Underwater landslides set off by the quake may have made the tsunami worse. Mudslides on land also swept into buildings.

FRANK FICHTMÜLLER/GETTY

finding on 1 October in a circular from the International Astronomical Union.

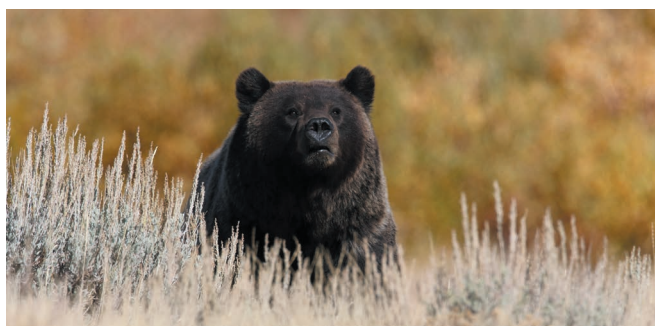
ENVIRONMENT

Sea-ice minimum

Arctic sea-ice cover following this summer's melt was the sixth lowest on record, the US National Snow and Ice Data Center in Boulder, Colorado, announced on 27 September. Sea-ice extent bottomed out for the season at 4.59 million square kilometres, tying with levels seen in 2008 and 2010. The sea ice reached its minimum this year on 23 September, which is one of the latest dates observed for this annual occurrence. The most recent calculations, from 2017, show that the September Arctic sea-ice extent has been declining by 13.2% per decade. A relatively cool July helped to slow this year's rate of loss. Since satellite records began in 1979, the 12 lowest extents have all happened in the past 12 years. The record low came in 2012, at 3.39 million square kilometres.

Yellowstone bears

A US federal court has restored endangered-species protections to grizzly bears (*Ursus arctos horribilis*, pictured) living around Yellowstone National Park. The 24 September



ruling reverses a controversial 2017 decision by the US Fish and Wildlife Service (FWS) to remove legal protections for the roughly 700 grizzlies in the Greater Yellowstone population. Judge Dana Christensen determined that the FWS had failed to consider the effects that removing these bears from the endangered-species list would have on other grizzly-bear populations in the contiguous United States. The restored protections mean that trophy hunts in Wyoming and Idaho that had been scheduled for this autumn won't occur. The hunts could have killed up to 23 grizzly bears. It's unclear whether the government will appeal the court ruling.

PEOPLE

CERN suspension

Europe's particle-physics laboratory, CERN, has suspended an Italian

theoretical physicist after he allegedly denied that physics suffers from a sexist bias and criticized positive-discrimination policies during a presentation at the lab. CERN, near Geneva, Switzerland, announced on 1 October that the physicist, Alessandro Strumia of the University of Pisa in Italy, was barred "from any activity at CERN with immediate effect, pending investigation into last week's event". Strumia gave his talk on 28 September. He disputes the characterization on social media of his presentation as sexist. In its statement, CERN said that Strumia's remarks were antithetical to its code of conduct and to its values. The University of Pisa and the European Research Council, which funds his research, also say that they are opening investigations. Strumia gave his talk at the lab's first

Workshop on High Energy Theory and Gender, in front of an audience largely made up of women. The reaction on social media was quick and fierce. In response, Strumia says: "I trust that the honest majority will understand that it is the truth, and that it was worthwhile to suffer such lynching for not submitting to censorship." Strumia, who was an 'invited scientist' at CERN, told *Nature* he hoped CERN "will want to talk and tell me what it was about my talk that was illegal."

FUNDING

Open access

Finland's national research funder has signed up to Plan S — a push by a group of European organizations to expedite the goal of making the results of research openly available. The Academy of Finland, which announced its move on 24 September, is the first organization to sign up since Plan S was launched by 11 funders last month. The now 12-strong coalition demands that, from 2020, papers resulting from the research they fund are immediately free to read on publication. The Finnish academy handed out more than €440 million (US\$510 million) in research funding in 2017.

SOURCE: CDC

TREND WATCH

The number of babies born in the United States with syphilis reached a 20-year high in 2017, according to a report by the Centers for Disease Control and Prevention (CDC).

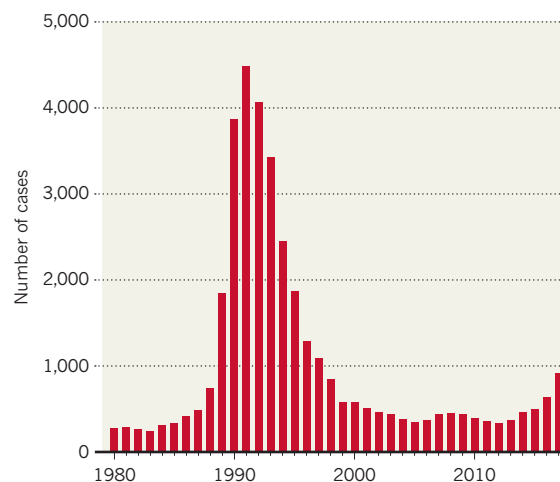
Reported cases totalled 918 last year, more than double the number four years earlier and the most cases in a single year since 1997. Before 2013, the number of cases had been dropping for 5 years. Syphilis is a sexually transmitted infection caused by the bacterium *Treponema pallidum*. Pregnant women who are infected with the bacterium can transfer it to their babies either through the placenta or

during birth. If the disease isn't treated, it can lead to premature birth, miscarriage, stillbirth or death of the baby soon after birth. Babies that survive might have deformed bones, jaundiced skin, or brain or nerve damage.

The rise in cases of syphilis in newborns — known as congenital syphilis — might, in part, be a result of physicians missing opportunities to screen and treat pregnant women with syphilis, says Virginia Bowen, an epidemiologist with the CDC in Atlanta, Georgia. "Many women who are giving birth to babies are not receiving timely prenatal care," says Bowen.

NEWBORN SYPHILIS ON THE RISE

The number of babies being born with syphilis has risen to a 20-year high in the United States.

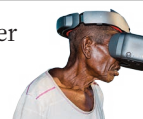


NEWS IN FOCUS

POLICY United States government launches review of fetal-tissue research **p.16**

ECOLOGY First plan emerges to clean up Amazon oil pollution in Peru **p.18**

NOBEL PRIZES Wins for laser wizardry and cancer immunotherapy **p.20**



COMMUNITY Researchers are increasingly partnering with non-scientists **p.24**

CHEN BIN/XINHUA VIA ZUMA



An agricultural official from Namibia learns about technology to combat desertification at a Chinese lab.

DEVELOPMENT

Science benefits in China's \$60-billion Africa splurge

Critics worry the investment will make African countries too reliant on an outside power.

BY DAVID CYRANOSKI

China wants to train Africa's next generation of scientists. Its lofty goal is to improve African science in fields from agriculture and climate change to quantum physics and artificial intelligence.

The training is one element of a much larger plan adopted by Chinese and African leaders at the third Summit of the Forum on China-Africa Cooperation in Beijing on 3-4 September. Chinese President Xi Jinping has

pledged US\$50 billion in grants and loans for infrastructure projects, medical programmes, clean-energy initiatives and other projects in Africa, and Chinese companies will invest another \$10 billion. The amount dedicated to training scientists is not known.

But some policy experts and scientists worry that African nations might become too reliant on other countries to provide training. Others doubt that the initiatives will truly boost African science, because similar projects planned at past forums have yet to produce noticeable benefits.

Few details have been released about how the money will be distributed among countries. The division is likely to be controversial, says Lina Benabdallah, who studies Chinese foreign policy in Africa at Wake Forest University in Winston-Salem, North Carolina. "It will be up to African leaders, political elites and their constituents to press for specific programmes to happen," she says.

Training is a pillar of the new plan. China will offer 50,000 scholarships for African people, including scientists, to study in ►

China, and will provide short-term training opportunities for another 50,000 people to travel to seminars and workshops.

The action plan also offers scholarships for postgraduate training in China and at African institutions, such as the Sino-Africa Joint Research Centre at the Jomo Kenyatta University of Agriculture and Technology in Juja, Kenya. The centre, which opened in 2013, collaborates with Wuhan Botanical Garden in China, and has produced dozens of academic papers in fields including biodiversity and climate-change monitoring.

China will also support a major expansion of the University of Health and Allied Sciences, a modern biomedical training institution in Ho, Ghana, which received \$20 million from the country in 2015.

“Developing indigenous talents locally is extremely important to the future of science in Africa,” says Tommy Karikari, a neurology researcher from Ghana who works at the University of Gothenburg in Sweden. The latest plan will dramatically expand training opportunities for African scientists, he says.

Karikari says that local scholarships and training facilities are important to ensure that some researchers stay in Africa. Many people currently train abroad because of a lack of opportunities on the continent, says Karikari. “It is expensive, and many beneficiaries do not return home, which affects the pool of trained scientists in Africa,” he says.

Benabdallah says the summit focused

particularly on ways to include African scientists in China’s global-diplomacy programme, the Belt and Road initiative. For example, the plan encourages researchers in Africa to join the Young Scientists Exchange Program, which pays for scientists to study in China for up to a year.

China has also promised to help countries develop real-world applications in quantum

“Developing indigenous talents locally is extremely important to the future of science in Africa.”

physics and artificial intelligence. But Benabdallah says there is a risk that African nations might become too dependent on other countries to provide training and skills. It is important

for African nations to be producers of science and technology, not just consumers, she says.

The plan also reaffirms China’s decades-long commitment to help improve agricultural science and practices and environmental protection in Africa. Analysts characterize this investment as a mix of profit-seeking, philanthropy and food security, as China seeks grains and oilseeds that it can bring back home.

The plan calls for new centres for joint research in environmental issues and geoscience, although their locations are yet to be announced. Other programmes will focus on safeguarding biodiversity and combating climate change and desertification. Five hundred senior agriculture experts from China

will also be sent to Africa to help modernize agricultural practices.

But Ademola Adenle, who studies sustainable development at Colorado State University in Fort Collins, is sceptical about China’s intentions. He says little knowledge has been gained from the more than 20 Chinese-government-funded agricultural-technology development centres created in Africa since 2006. The centres lack transparency and mainly represent Chinese commercial interests, he says. One of them reportedly sells farm equipment, mushroom powder and dried mushrooms to local people.

“Since this initiative kicked off, I am not aware of any significant breakthrough in agriculture research and development, or any type of innovation that could transform agricultural development,” he says.

China’s agriculture ministry did not respond to questions about the agricultural-technology centres by *Nature’s* deadline.

Adenle hopes that the forum will result in training for agricultural scientists to improve local farming techniques. But if these initiatives just give China more access to Africa’s natural resources, it could spell doom for the continent, he says.

For China’s investments to help Africans harness science and technology, there will need to be more public discussion of the trade agreements and political deals as they’re worked out. “There is no doubt that China has invested a lot of money in Africa,” says Adenle. “But we need more transparency.” ■

POLICY

Fetal-tissue work under scrutiny

US government will examine federally funded studies.

BY SARA REARDON

The US government has begun a review of federally funded studies that use fetal tissue, a move that critics fear could be a first step toward curbing such research.

Following complaints from anti-abortion groups and Republican lawmakers, the Department of Health and Human Services (HHS) plans to evaluate “all research involving fetal tissue” and “all acquisitions involving human fetal tissue”. In a statement on 24 September, the department also said that it had cancelled a US\$15,900 contract between the Food and Drug Administration, which it oversees, and Advanced Bioscience Resources (ABR), a non-profit tissue supplier in Alameda, California.

According to the contract, which the FDA awarded in July, agency researchers would implant the human fetal tissue provided by the company into mice that lacked immune systems. The goal was to give the animals human-like immune systems and use them to evaluate the safety and efficacy of various drugs.

The HHS said that it cancelled the contract because it “was not sufficiently assured that the contract included the appropriate protections applicable to fetal tissue research or met all other procurement requirements”.

The action comes after 85 members of the US House of Representatives sent a letter to FDA commissioner Scott Gottlieb on 17 September, claiming that ABR might have violated federal law by selling “the body parts of

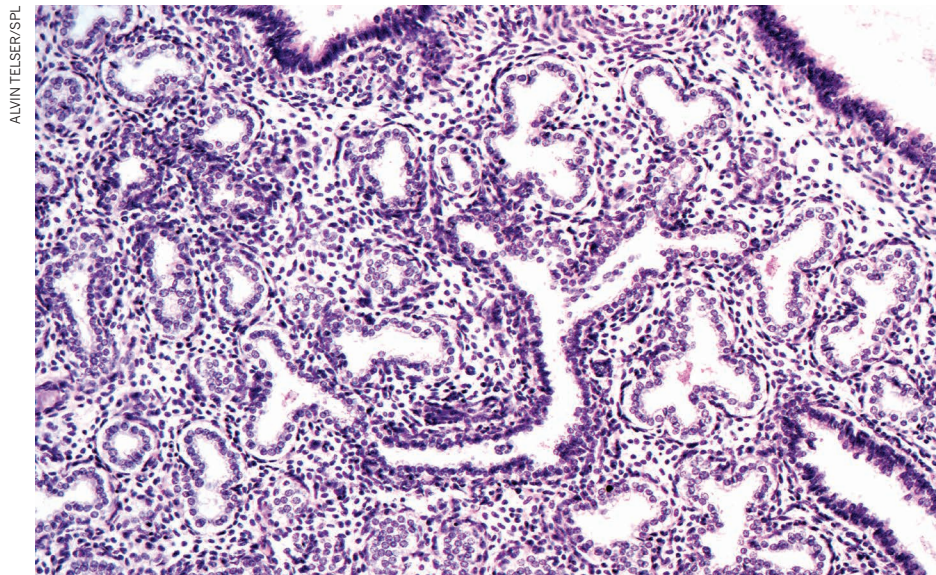
children” for a profit. In 2016, a special House committee — the Republican-led Select Investigative Panel on Infant Lives — had referred the company to the US Department of Justice for investigation. ABR did not immediately respond to *Nature’s* requests for comment.

“We are alarmed that the FDA has continued to award contracts to ABR for the procurement of human fetal tissue,” the lawmakers wrote to Gottlieb.

The HHS has offered little detail about its review of fetal-tissue contracts and research. In its statement, the department said that it is auditing “all acquisitions involving human fetal tissue” to ensure that firms that supply the tissue adhere to federal regulations.

The department is also reviewing research involving fetal tissue “to ensure the adequacy of procedures and oversight of this research in light of the serious regulatory, moral, and ethical considerations involved”, as well as whether alternatives to such research exist. An HHS spokesperson declined to comment on how long the process would take, but said that it would encompass research funded by the department.

That decision has “troubled” the International Society for Stem Cell Research in Skokie, Illinois, the group said in a statement on 27 September. “The directive appears to come after



The US government is reviewing research that involves fetal tissue, such as these lung cells.

intensive lobbying efforts by special interest groups with the goals of delaying or curtailing scientific research,” the group wrote. “Research that has saved lives, and will likely save more,

should not be delayed for political reasons.”

Larry Goldstein, a neuroscientist at the University of California, San Diego, says that it’s hard to know why the HHS decided to

cancel the contract. “I think the question is whether there’s an attempt to politicize this or whether we can keep to straight scientific and medical merit,” he says.

Goldstein is concerned that a ban or heavy restrictions on federally funded experiments with fetal tissue could harm research, particularly on human development, organ regeneration and determining whether tissue created from stem cells recapitulates the real thing.

Moreover, he says, the fetal tissue used in research would otherwise be discarded. “Scientists are simply asking, if you’re going to throw the tissue away anyway, can you at least donate it to important medical research?”

Renate Myles, a spokesperson for the US National Institutes of Health (NIH), which sits within the HHS, says that the agency does not have any standing contracts with any providers of human fetal tissue. “We agree that it is important that research involving human fetal tissue should be consistent with the statutes and regulation governing such research, and reminded NIH-funded institutions that awards are conditioned upon compliance of all applicable federal, state and local laws and regulations,” she says. ■

PALAEONTOLOGY

Early-Jurassic dinosaur find puts evolution of walking to test

Fossils suggest quadrupedalism evolved 10 million years earlier than thought.

BY SARAH WILD

Researchers have discovered fossils from South Africa’s largest dinosaur yet — a find that they say changes their understanding of how four-legged walking evolved in a lineage that includes some of the biggest animals ever to have walked the planet.

The newly described species, *Ledumahadi mafube*, would have weighed about 12 tonnes, and is a type of sauropodomorph: a large group of dinosaurs with long necks and tails. When *L. mafube* lived, around 200 million years ago during the early Jurassic period, it would have been the largest animal walking on Earth (B. W. McPhee *et al.* *Curr. Biol.* <http://doi.org/cvbd>; 2018).

Palaeontologist James Kitching first found fossils of *L. mafube* in 1988 near South Africa’s border with Lesotho. But the bones were left on a shelf for more than a decade and ‘rediscovered’ only in the 2000s, in the collections of the University of the Witwatersrand in Johannesburg, South Africa. Palaeontologists returned to the site in

2010 and completed the excavation last year.

Most of South Africa’s dinosaur discoveries have been of animals that would have weighed about five tonnes, says study co-author Jonah Choiniere, a palaeoscientist at Witwatersrand. The discovery of such a heavy creature shows “we don’t know the dinosaurs of South Africa as well as we thought”, says Choiniere. Other researchers agree that *L. mafube* was probably the largest animal of the early Jurassic.

“We don’t know the dinosaurs of South Africa as well as we thought.”

A WALKING GIANT

But the find is also significant because it seems to show that quadrupedalism — walking on four legs — emerged in this lineage of dinosaurs at least 10 million years earlier than thought, and then disappeared before returning.

Researchers knew that other, later sauropodomorphs, such as *Brontosaurus*, had straight, ‘columnar’ limbs that could support their huge

mass, often in the region of 80 tonnes. They also knew that some sauropodomorphs that came before *Brontosaurus* and its kind, but after the time of *L. mafube*, walked on two legs. “We thought [quadrupedalism] might be a one-time evolution: a quadruped walks once, is successful, and it sticks in that lineage,” says Choiniere. But the latest proposal — which is based on a ratio of the circumferences of thigh and arm bones, calculated from dinosaur specimens and animals alive today — that *L. mafube* also walked on four legs changes that view. The finding hints at evolutionary experimentation: some sauropodomorphs had quadrupedalism and then the group lost it, says Choiniere.

That claim is controversial, says Michael Benton, a palaeontologist at the University of Bristol, UK. Unlike later sauropods, *L. mafube*’s legs flexed out to the sides, a stance that is typically able to carry less mass than columnar limbs. “What’s needed next is a true biomechanics test of whether 12 tonnes is the maximum size an animal can reach without having columnar limbs,” he says. ■

ECOLOGY

Peru plans oil clean-up

Proposal would address decades of pollution in the country's largest Amazon oil field.

BY BARBARA FRASER

Nearly half a century after poorly regulated oil producers began dumping billions of litres of wastewater and other toxic substances into the rivers and tropical forests of northern Peru, the government is taking its first steps towards cleaning up the damage in the country's oldest and largest Amazonian oil field.

Government contractors are drawing up plans for the remediation of 32 polluted sites — out of up to 2,000 identified so far — in an oil-producing area known as Block 192. The sites have been listed as priorities by environmental authorities and local indigenous organizations. Meanwhile, a government-funded study conducted by the United Nations Development Programme (UNDP), and released publicly in August, recommends a more comprehensive remediation strategy based on health risks, along with changes to environmental regulations.

Block 192 covers about 4,970 square kilometres overlapping the Pastaza, Corrientes and Tigre rivers, which flow into the Marañón, one of the main tributaries of the Amazon. It is located in a remote region near Peru's northern border with Ecuador that is inhabited mainly by Quechua, Achuar and Kichwa people. The US-based company Occidental Petroleum began operating there in the mid-1970s, two decades before Peru passed environmental regulations for the oil and gas industry. Oil production in Block 192 peaked in 1982, at 120,000 barrels a day.

For four decades, the environmental effects of this activity were largely uncontrolled and unremediated. The hot, salty, metal-laden water pumped out of wells with oil was dumped into streams and rivers until 2009, nine years after the Argentinian firm Pluspetrol took over Occidental's lease. In 2015, Occidental announced that it had settled — for an undisclosed sum — with five Achuar communities who sued in a US court over pollution in



Segundo Cariajano Hualinga, president of a Kichwa community in Peru, stands on an oil pipeline.

the Corrientes River. The company denies that its operations contaminated Block 192.

The discharge of polluted water, combined with ongoing spills from the pipelines crisscrossing the block, resulted in chronic exposure of fish, frogs and other aquatic life to salts, heavy metals and hydrocarbons, according to the UNDP-sponsored study. Residents of villages in Block 192 have also been exposed to these pollutants. Until 2015, when the government began installing temporary water-treatment plants, rivers and streams were the only source of water for drinking, cooking and washing. Villagers recall bathing in salty water and pushing oily scum aside to draw water for drinking or cooking.

Many continue to question whether fish are safe to eat. Near the Kichwa community of Doce de Octubre on the Tigre River, oil seeps to the surface of a former lake in the midday heat and a small stream of salty water trickles past decades-old pipes, says Segundo Cariajano Hualinga, the community's president.

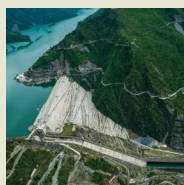
Indigenous federations have fought for more than a decade to get Peru's government to commit to remediating sites polluted with oil, and to agree to the UNDP study. Plans are under way to clean up the 32 priority sites, but a US\$15-million contingency fund is nearly exhausted. The UNDP study estimated that remediating 92 sites could cost \$300 million. Although full funding is not yet assured, Peru's minister of energy and mines said on 6 September that he would request an extra \$51.5 million in 2019 to clean up Block 192.

ECOSYSTEM OBSTACLES

But the Amazon itself presents major challenges to any clean-up effort. Environmental remediation is daunting in rainforest ecosystems, with their seasonal flooding, varying water chemistry and poorly understood groundwater flows, says Margarita Núñez, a biologist in Costa Rica who headed the UNDP study team. The saturated, nutrient-poor clay soil in Block 192 contains little oxygen. She



TOP NEWS



Landslides threaten Himalayan hydropower dream
go.nature.com/2lmutub

MORE NEWS

- 'Severe' figure manipulation found in studies from plant lab go.nature.com/2n7eiap
- Syphilis cases in US newborns spike to 20-year high go.nature.com/2zjautm
- France calls for slim 2% increase to research budget go.nature.com/2p3rzxe

NATURE PODCAST



Targeting latent HIV; bird personalities; and update on Hayabusa2 mission nature.com/nature/podcast

BARBARA FRASER

PETE MCBRIDE/GETTY

says that this makes it difficult to use one of the cheapest, least invasive clean-up methods — introducing microorganisms that can break down oil and gas pollution.

Because of those conditions, along with the accumulation of metals from decades of polluted-water discharge, the UNDP study recommends a combination of measures. These include bioremediation with plants and microbes, removal and incineration of contaminants, stabilization or solidification of polluted areas, and thermal desorption, in which heat is used to separate individual contaminants from a mixture.

Many clean-up options put forth by the study would be difficult to use in Block 192, says Raúl Yusta García, a chemical engineer at the Monterrey Institute of Technology in Mexico. His research suggests that pollution levels in water discharged from oil operations have been underestimated.

Because Peruvian law lacks a clear definition of environmental remediation, government officials, company executives and local communities have different expectations about what a clean-up effort should entail. Although the “occidental, technified world” sees the goal of remediation as reducing health risks, Indigenous people think of it “more as restoration”, returning a site to its former state, says Fernando Morales, an environmental chemist at Simón Bolívar University in Caracas, who worked on the UNDP study.

Peru's environmental standards are inadequate for gauging risks to human health and the environment, especially the aquatic ecosystems on which people depend for food, says Diana Papoulias, an aquatic toxicologist retired from the US Geological Survey, who was also on the UNDP team.

The country has quality standards for water and soil, but not for sediments — where hydrocarbon residue and metals have probably settled, and which might be redistributed throughout the forest by floodwaters during the annual rainy season, she says. The standards that do exist were taken from nations such as the United States, Canada and the Netherlands, whose largely temperate ecosystems are very different from the Peruvian Amazon.

That very ecosystem might put local residents at further risk from pollution. Health exams have found cadmium and lead in Indigenous villagers' blood and urine. Núñez and her colleagues suspect that people might be more likely to absorb those elements in nutrient-poor environments that lack sufficient calcium and magnesium.

The UNDP team has recommended further study of the ecosystem, but that should not delay the clean-up, Núñez says. “Of course you'd do a better job of remediation if you had more knowledge, but I believe the remediation, as such, has to begin now.” ■

EQUALITY

Nobel committees to tackle gender skew

Nominators will be asked to consider diversity in gender.

BY ELIZABETH GIBNEY

When Donna Strickland won a Nobel Prize in Physics this week (see page 20), she was the first female winner in more than 50 years. Over that period, just one woman has won in chemistry (*Nature* went to press before this year's prize was announced).

This gender imbalance is the subject of increasing criticism, much of which is aimed at the Nobel committees that award the honours. In the awards' history, women have won only 3% of the science prizes (see ‘Nobel imbalance’), and the overwhelming majority of prizes have gone to scientists in Western nations. Some argue simply that the prizes tend to recognize work from an era when the representation of women and non-Western researchers in science was even lower than it is today. But studies repeatedly show that systemic biases remain in the sciences — and the slow pace of progress was especially evident in 2017, when there were no female laureates for the second year in a row.

Göran Hansson, secretary-general of the Royal Swedish Academy of Sciences, says that

even now, too few women might be nominated.

For the first time, the committees will explicitly call on nominators to consider diversity in gender, geography and topic for the 2019 prizes. The request will be included in invitation letters scheduled to go out this month to the thousands of scientists asked to nominate candidates for each prize. “We need the scientific community to see the women scientists, and to nominate those who have made outstanding contributions,” he says.

“The smallest possible nudge can make a difference, so I praise them for that,” says Curt Rice, president of Oslo Metropolitan University and head of Norway's Committee for Gender Balance and Diversity in Research. Other measures and suggestions are in the works to improve gender balance, including changes to nomination committees and nomination rules.

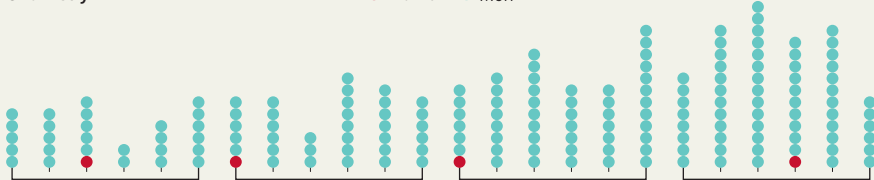
Hansson says that the diversity measures are not about improving the statistics, but about helping the best scientists to win by ensuring that outstanding women are not overlooked. “We are admittedly slow, but we are aware of the situation and we work on it,” he says. ■

NOBEL IMBALANCE

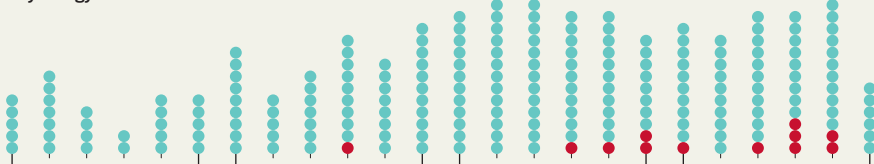
Of the 604 Nobel medals awarded in scientific disciplines, just 19 have gone to women.

Chemistry

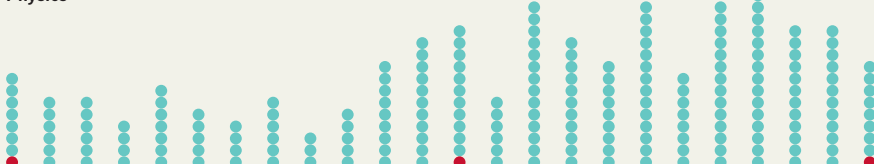
● Women ● Men



Physiology or Medicine



Physics



1901–30

1931–60

1961–90

1991–2018

One stack represents 5 years, except for 2016–18 (2016–17 for chemistry).



Members of the Nobel committee for physics announce this year's prize.

Nobel Prize in Physics; the last female scientist to win it was Maria Goeppert Mayer, in 1963. "I don't know what to say, I'm honoured to be one of those women," said Strickland. Göran K. Hansson, secretary general of the Royal Swedish Academy of Sciences in Stockholm, said the academy is "taking measures" to encourage more nominations of female scientists, but that those measures did not affect this year's prize.

HANNAH FRANZEN/AFP/GETTY

POWERFUL PULSES

Short-lived laser pulses allow scientists to spy on processes that are over in a heartbeat. But before Strickland and Mourou's technique, the intensity of such pulses was limited because the high power risked destroying the amplifier needed to create them. The pair's breakthrough was to stretch out a laser pulse in time. This reduced the power of the light and made it possible to use conventional amplifiers, before the pulse is squeezed back together. Because fleeting pulses cause less damage, they have also found uses in laser eye surgery. "Nobels are awarded for a discovery or an invention. This really bridges the two," says John Dudley, an optical physicist at the University of Franche-Comté in Besançon, France.

At 96, Ashkin is the oldest-ever Nobel laureate. His prize-winning work began immediately after the laser's invention, in 1960. Lasers exert a gentle pressure, which Ashkin realized could be used to manipulate tiny objects without damaging them. His experiments with micrometre-sized spheres showed that the particles were drawn to the highest-intensity region in a beam of light. This led to a way to sculpt lasers to trap, levitate and move objects. Ashkin discovered that these 'optical tweezers' could capture bacteria, viruses and living cells. Miles Padgett, an optical physicist at the University of Glasgow, UK, says that the impact of Ashkin's work has been universally recognized. Today, optical tweezers are used in myriad applications, from separating healthy blood cells from infected ones to engineering nanoscale materials. ■

PRIZES

Laser tricks win physics Nobel

Laureates include first female winner in 55 years.

BY DAVIDE CASTELVECCHI,
ELIZABETH GIBNEY & MATTHEW WARREN

A trio of laser scientists has won the 2018 Nobel Prize in Physics for work using intense beams to capture superfast processes and to manipulate tiny objects. The laureates include Donna Strickland, who is the first woman to win the award in 55 years.

Strickland, at the University of Waterloo, Canada, shares half of the prize, worth 9 million Swedish krona (US\$1 million), with her former supervisor, Gérard Mourou, now at the École Polytechnique in Paris. Arthur Ashkin, at Bell Laboratories in Holmdel, New

Jersey, won the other half of the prize.

Strickland and Mourou pioneered a way to produce the shortest, most-intense pulses of light ever created. These are now used throughout science to unravel processes that previously appeared instantaneous, such as the motion of electrons within atoms. Ashkin won the prize for developing 'optical tweezers', beams of laser light that can grab and control microscopic objects such as viruses and cells.

"First of all, you have to think it's crazy, so that was my first thought," said Strickland during the announcement of the prizes on 2 October. "And you do always wonder if it's real."

Strickland is the third woman ever to win the

AWARDS

Cancer immunologists scoop medicine Nobel prize

One of the hottest areas in cancer research, immunotherapy can dramatically extend lives.

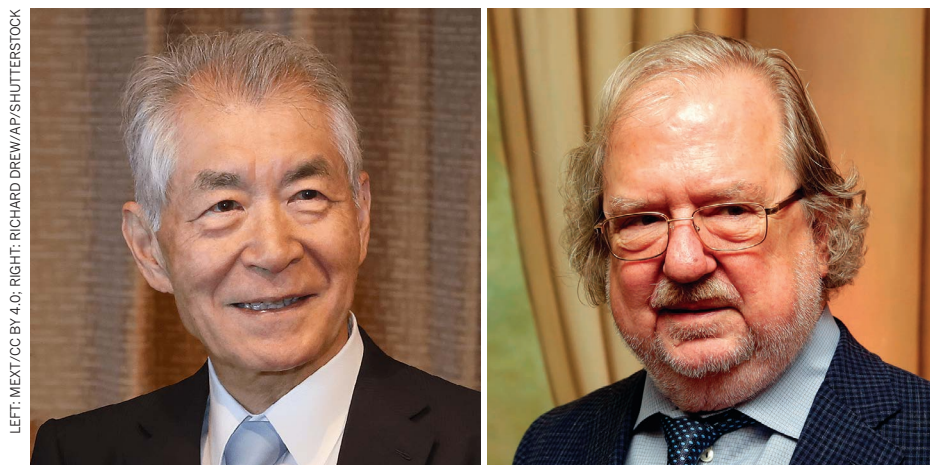
BY HEIDI LEDFORD, HOLLY ELSE AND
MATTHEW WARREN

Two scientists who pioneered a new way to treat cancer have won the 2018 Nobel Prize in Physiology or Medicine. James Allison at the University of Texas MD Anderson

Cancer Center in Houston and Tasuku Honjo at Kyoto University in Japan showed how proteins on immune cells can be used to manipulate the immune system so that it attacks cancer cells. The approach has led to therapies that have extended lives, and even wiped out all signs of disease in some people with advanced cancers.

"To have my work really impact people is one of the best things I could think about," said Allison at a press conference on 1 October, the day the 9-million-Swedish-krona (US\$1-million) prize was announced. "It's everybody's dream."

In the 1990s, Allison, then at the University of California, Berkeley, studied a protein,



Tasuku Honjo (left) and James Allison share the 2018 Nobel Prize in Physiology or Medicine.

CTLA-4, that acts as a brake on immune cells called T cells. In 1997, he and his colleagues engineered an antibody that binds to CTLA-4, unleashing T cells to attack cancer in mice. A clinical study in 2010 found that the antibody had a striking effect on people with advanced melanoma, a form of skin cancer¹.

Working independently of Allison, in 1992, Honjo discovered the T-cell protein PD-1, which acts as a brake on the immune system by a different mechanism. Research showed the protein to be highly effective against several

human cancers, including lung cancer². Some people with metastatic cancer went into long-term remission, raising the possibility of a cure.

Clinical work on 'immune checkpoint therapy' has since developed apace. Treatments that block PD-1 have proved to be effective in lung and renal cancers, lymphomas and melanoma. And combined therapies that target CTLA-4 and PD-1 in people with melanoma showed that this approach can be even more effective than CTLA-4 alone³. Trials are now under way to evaluate the efficacy of checkpoint therapy

against most types of cancer, and scientists are looking for other checkpoint protein targets.

Others also made important early discoveries about checkpoint inhibitors, notes Gordon Freeman, an immunologist at the Dana-Farber Cancer Institute in Boston, Massachusetts, who was disappointed not to be recognized. Freeman, along with immunologists Arlene Sharpe at Harvard Medical School in Boston and Lieping Chen at Yale University in New Haven, Connecticut, also studied checkpoint proteins, and a molecule that binds to PD-1, PD-L1. The US Food and Drug Administration has since approved drugs that target PD-1 and PD-L1.

But immunologist Jerome Galon of the French national biomedical research agency, INSERM, in Paris, describes Honjo and Allison as "the obvious two first choices".

The immune approach to fighting cancer has come a long way. Allison met with considerable resistance when he first tried to get pharmaceutical companies interested. But in 2012, when PD-1 inhibitors were shown to work against lung cancer, the field ignited.

It's been thrilling to watch the field develop, says Freeman. "It's wonderful because so many cancer patients are doing better." ■

1. Hodi, F. S. *et al.* *N. Engl. J. Med.* **363**, 711–723 (2010).
2. Topalian, S. L. *et al.* *N. Engl. J. Med.* **366**, 2443–2454 (2012).
3. Larkin, J. *et al.* *N. Engl. J. Med.* **373**, 23–34 (2015).



Science shared

Those who were once the subjects of scientific enquiry are increasingly in the driver's seat. A special issue explores the co-production of research.

GARTH CRIPPS

From people with HIV selecting which trials of antiviral therapies get funded, to farmers of smallholdings guiding weather monitoring, the people affected by research are increasingly getting involved in it. They are shaping how projects are conceived, supported, done, assessed, disseminated and rated. They are partners in research production.

This special issue looks at the promise and the pitfalls of co-production for the stakeholders, scientists and societies now working shoulder to shoulder. As one advocate describes it: "It's about getting everybody round the table so you're valuing the knowledge everybody has."

A series of case studies on page 24 illustrates the many forms such research can take. They include a public-health researcher who has been working to curb childhood obesity with members of the Osage Nation, a Native American community in Oklahoma; and climate modellers embedded with city planners in nine southern African cities to help determine the research and infrastructure needed to adapt to climate change. The stories highlight common themes: co-production takes people out of their comfort zones, but the pay-off comes in the form of enhanced trust

and communication. Importantly, the research has a much better chance of making a difference to the people involved.

Those who were previously outside the academic system are also becoming gatekeepers for research: helping to decide what gets funded, published and evaluated. A collection of Comment articles describes how patients and carers are invited to review manuscripts at *The BMJ* (see page 30) and grant applications at the California Institute for Regenerative Medicine (see page 31). In some cases, they encourage risk-taking, in others, they rein in false hope. Another article calls for the wider use of co-created evaluation tools to improve and incentivize research co-production (see page 32).

So how do you join the revolution? Public-involvement manager Gary Hickey offers five principles for co-producing research on page 29. Chief among these is to share power. But, as he writes, co-production won't happen just because it is a good thing: research partners need to change their practices and cultures. Getting everybody around the table is worthwhile, but it takes work. ■



CO-PRODUCTION OF RESEARCH
A Nature special issue
[nature.com/collections/coproduction](https://www.nature.com/collections/coproduction)

PARTNERS IN SCIENCE

The people who should benefit from research are increasingly shaping how it is done.

BY CASSANDRA WILLYARD, MEGAN SCUDELLARI AND LINDA NORDLING

Valarie Blue Bird Jernigan knew she had to tweak some standard scientific practices when she started her latest research project. One of the first things to go was the usual concept of a control group — people who would not receive interventions to encourage healthy eating. That wouldn't be fair to the people of the Osage Nation, a Native American people in northeastern Oklahoma.

Another concept to ditch was the idea that she was studying a group at all. Jernigan, a public-health researcher, who is Native American herself, has treated the Osage people as equal partners from the first day of the project. It took two years and seemingly endless rounds of community discussions to get the study off the ground, but Jernigan wouldn't have had it any other way. This kind of research “isn't just about proving your hypothesis,” she says. It's more about improving people's lives and, at the same time, helping them gain the skills to do science.

Jernigan's approach, often referred to as community-based participatory research, has been gaining traction for the past two decades. It has become particularly important for research that involves indigenous and other populations who have been mistreated by scientists in the past. The Havasupai tribe in Arizona, for example, waged a lengthy legal battle with Arizona State University in Phoenix over researchers' misuse of blood samples that the tribe had provided for a diabetes study in the 1990s. The samples were eventually returned as part of a settlement two decades later. The lessons learnt from the event have set the tone for how best to do research involving Native Americans.

Community participation has become the norm. “In minority communities, it's probably the primary research methodology,” says public-health researcher Alexandra Adams, director of the Center for American Indian and Rural Health Equity at Montana State University in Bozeman. “It reduces mistrust, it improves dissemination and it improves cooperation.” The goal of such efforts is the co-production of research, in which the stakeholders who are supposed to benefit from a strand of research become active partners in conducting it. Scientists from disciplines as varied as archaeology, public health and climate change have embraced the approach, working with community members on many different aspects, from formulating study questions and design, to doing experiments and analysing and reporting results.

Nature talked to three groups that have built successful co-produced projects. Their experiences reveal the challenges and rewards that come with the open and collaborative exchange of ideas. The work veers away from the standard outputs of science, such as talks and papers, and expands the idea of what it means to be a scientist and a collaborator.



A PLACE AT THE TABLE

Jernigan's latest project with the Osage people wasn't wholly her idea. It started with Raymond Red Corn. As a child growing up in the Osage Nation, Red Corn helped his parents to harvest the dusky red ears of maize (corn) and process them into corn soup and hominy, a food made from soaking kernels in lye or wood ash until they go puffy. Taking the maize from seed to soup is something the Osage have done for centuries. But that tradition has nearly disappeared. “I couldn't hardly find anyone younger than me that had ever done it, even in the most traditional families,” he says.

Four years ago, Red Corn was elected assistant chief of the Osage Nation. Right away, he started looking for a spot to plant traditional maize and other crops. Fresh fruit and vegetables are hard to come by in Osage County. Since the 1970s, the Osage people have increasingly relied on canned and processed foods that are high in salt, fat and sugar.

Red Corn wants to see the community take

GARTH CRIPPS



Researchers excavate an ancient cattle pen for the Morombe Archaeological Project.

mistrustful of the scientific enterprise. In the past, investigators have used tribal members as unwitting

participants in unethical and dangerous experiments. And, as in the Havasupai case, scientists have at times withheld information from the communities they have studied and largely ignored tribal concerns.

When Native Americans think of health studies, they often think of “helicopter researchers”, Jernigan says — scientists who fly in, collect data and blood samples, and then leave. “And they never see one benefit.” What’s more, working with indigenous communities means dealing with sovereign governments, some of which have their own institutional review boards. “You have to go through all these extra layers of protections,” Jernigan says. These days, collaboration and co-production aren’t just ethical, they are mandatory. “There’s almost no other way of doing it,” she says.

As a first step, Jernigan proposed launching a pilot study to work out what the community actually wanted. The team surveyed everyone from community members to leadership, and found that people seemed most interested in the idea of community gardening. They wanted to use locally grown crops to help supply some of the tribally run programmes for children and older people.

But boosting the supply of fresh fruit and vegetables is only half the battle; people also wanted to increase the desire for healthy foods. So Jernigan worked with the Osage to design a community programme aimed at getting young children and their families to eat more fruit and vegetables. The trial, called Food Resource Equity and Sustainability for Health, or FRESH, launched in January. The team came up with new, healthier menus for a programme that provides care for children aged 3–5 from low-income backgrounds. The researchers also provided the schools with demonstration gardens. Each week, the teachers spend 90 minutes telling stories about food, working with the children in the garden, and conducting a simple cooking lesson. On Fridays, the children take home a healthy meal kit to prepare with their families. Meanwhile, their parents take part in a 15-week online workshop.

The cultural elements are important. Parents are encouraged to attend a monthly family night, where they talk about foods they remember eating when they were young, what they eat now, where it comes from and why they choose certain foods. Jernigan’s team has given video cameras to families to record their own food stories. “There’s a lot of realization about ►

“YOU NEVER ASK SOMETHING OF SOMEONE WITHOUT GIVING THEM SOMETHING BACK.”

back control of its food supply. By restoring their connection to the land and its lost food traditions, he thinks, they just might be able to rewind to a healthier lifestyle. The efforts might even help to tackle the high rates of obesity and diabetes in Native Americans in the area. In the Osage Nation, “everything we do revolves around food”, Red Corn says. “You can’t heal the community unless you heal the food system.”

Red Corn and other tribal leaders hoped that providing locally grown fresh foods would yield obvious health benefits, but they weren’t equipped to measure those benefits themselves. So, they reached out to Jernigan at the University of Oklahoma Health Sciences Center, who

is a member of the Choctaw Nation. Jernigan has spent the bulk of her career testing strategies to improve the food environment on reservations as a way to enhance health. She has another project with two other Native American communities in Oklahoma to get healthier foods into their convenience stores.

Research on marginalized groups can be fraught, and working with tribal communities is especially complicated. A history of research abuses has left many Native Americans



CO-PRODUCTION OF RESEARCH
A Nature special issue
[nature.com/collections/coproduction](https://www.nature.com/collections/coproduction)

► the connections between language, identity and indigeneity,” she says. “Those are the kinds of things that typical health-science interventions don’t address.” And “those are the kinds of things that offer really our best hope for health”.

The FRESH study will look at whether the interventions actually increase children’s consumption of fruit and vegetables, and their willingness to try them. The researchers will also look at health measures, such as body mass index and blood pressure, in the families.

Collaborative research isn’t always easy. Jernigan often has to strike a balance between scientific best practice and the community’s needs and desires. For example, the ideal way to test an intervention is often through a randomized trial. But with FRESH, the researchers couldn’t use a true control group. It wouldn’t be ethical to deny some of the participants the resources that the study provides. Instead, the team adopted a ‘wait-list control’ design. In the first phase of the project, two communities receive the intervention and two serve as the controls. When the 15-week intervention is complete, the control communities join the experimental arm. “That was a way to have a control group, but then to be able to tell the community that everybody gets the intervention,” Jernigan says.

Despite the challenges, Jernigan has no regrets. “I had seen traditional research in my training, and to me that seemed so myopic,” she says. And she was never interested in generating knowledge for its own sake. In traditional Native American culture, “you never ask something of someone without giving them something back”.

VALUE IN THE PAST

Kristina Douglass remembers sitting around a fire after dinner in Madagascar, listening to the conversations of women who spent their lives gathering shellfish among the sand and rocks of the Velondriake Marine Protected Area, a network of 25 fishing communities along the country’s southwest coast. The women were discussing a new variety of bivalve shellfish they had identified. “I couldn’t see the difference, but they insisted,” recalls Douglass, an archaeologist at Pennsylvania State University in University Park, who has been studying the fossilized remains of shellfish and other animals found in early human settlements in the area. Their conviction and their experience with the organisms suggest to her that they are probably correct.

Douglass directs the Morombe Archaeological Project (MAP), which is reconstructing the impact of human settlement on the Velondriake area, a biodiversity hotspot where pygmy hippopotamuses and giant tortoises once roamed. Since 2012, the project team has been conducting drone-assisted surveys, excavating fossils and preserving DNA. Team members have also been recording the oral histories of local elders, to explore the migration of clans in the area and preserve their histories.

Douglass is the only person on the team with a PhD. The others are members of the region’s five ancestral clans, and hail from three local communities: Vezo fishers, Masikoro farmers and herders, and Mikea foragers. Few have completed secondary school, and many cannot read or write. Yet Douglass considers them the experts, “and great field researchers”, she says.



Including locals in research was a priority for Douglass, even before MAP started. “I came to archaeology with a vivid sense of the absurdity,” she says. Relatively large sums of money, resources and time were going into studying people of the past while their descendants gained little from the findings. If the research didn’t have relevance for people living today, there wasn’t a point, she decided.

A region’s citizens can and should have a choice about how their area is represented and what research is valuable to them, says Eréndira Quintana Morales, an archaeologist at Rice University in Houston, Texas, and a collaborator with Douglass. In June she visited Andavadoaka, the fishing community where MAP is based, to teach a workshop on preparing and cataloguing the bones from contemporary fish to create a local reference collection. The intensive, co-production research effort exemplified by MAP also makes for better science, she says. “We go in with our own biases when we’re not open to learning from community members.”

The Velondriake women, for example, have local ways of describing taxonomic relationships between animals, such as referring to shellfish from different families as male and female versions of one another. Prompted by that knowledge, MAP now includes a project exploring how the classification system affects



Raymond Red Corn (centre) of the Osage Nation studies the impact of food sovereignty on his community.

RED CORN: NICK OXFORD FOR NATURE. MADAGASCAR: GARTH CRIPPS



A local historian in Madagascar describes live drone footage of an archaeological site.

Buckley of the Sea-Fisheries Protection Authority in Ireland. She leads a species-management

project with fishers along the coast of Kenya, and has listed non-scientists on papers. “There are certain people that I would like to make an author, but, according to criteria for a journal, I shouldn’t put them on.”

Douglass is planning several papers for the coming year in which she hopes to include the whole team as co-authors. “For me, it’s the next step in building this collaboration.” But it isn’t easy. Team members do not have regular Internet access, and Douglass lacks the funds to fly them to Pennsylvania to work in person. And such funds would be hard to obtain. Funders have already balked at her level of spending in Madagascar, with grant reviewers commenting that Douglass’s project budget for paying locals — about US\$200 per month, a living wage in the area — was unreasonably high.

Co-production research remains peripheral in archaeology, says Douglass. For the field to wholly embrace it, archaeologists will have to start questioning how they run their projects, she says. “There’s such a long and entrenched history of practising archaeology in a way that enforces certain power dynamics,” says Douglass. “It would make a lot of people uncomfortable to have to sit down and think about how you really collaborate.”

Plus, it takes time and effort, says Buckley. “It’s not just a science project. You need to approach people at their level and go back again and again and again.”

TEACHABLE MOMENTS

By September 2016, the residents of Zambia’s capital city, Lusaka, were getting desperate. The city of around 2 million people was withering in a drought. Maize harvests had dropped by about 20% from the year before, driving up food costs. And reduced water flow through the country’s main hydroelectric dam had triggered rolling blackouts in the region.

That month, Lusaka held its first ‘learning lab’, a gathering of city planners, policymakers and climate scientists intended to improve climate-related decision-making in the country. As the meeting got under way, people were looking to the researchers for answers. Chief among their concerns: when was it going to rain again?

But the scientists were not there to give answers. They were there to listen as part of the Future Resilience for African Cities and Lands (FRACTAL) programme, a co-production effort designed to improve the alignment of research and policymaking in nine southern African cities. ►

“IT’S NOT JUST A SCIENCE PROJECT. YOU NEED TO APPROACH PEOPLE AT THEIR LEVEL AND GO BACK AGAIN AND AGAIN.”

the ways in which locals choose to harvest species or leave them alone.

Leading the daily operations for many of the MAP projects is George ‘Bic’ Manahira, one of Douglass’s first full-time team members. Manahira, who is from Morombe and speaks five languages, began participating in MAP as a volunteer in 2012. He joined the staff a year later as the field manager. “I was curious how they do stuff,” says Manahira. “And I wanted to know my story.”

Thanks to the team’s efforts, that story is now being told. In a paper published this year, Douglass and the MAP team collected and analysed animal fossils from coastline rock shelters in the area (K. Douglass *et al. Quat. Int.* 471, 111–131; 2018). From roughly 1,400 years ago to the start of the twentieth century, settlers in the region harvested only certain marine species while leaving others untouched, and the team found no evidence linking humans to the extinction

of many large fauna, such as the giant tortoise. The finding challenges the assumption that rural communities are blanket consumers of the resources around them, which has implications for contemporary conservation efforts.

With the MAP team, Douglass asks members to participate in almost every aspect of research: experimental design, fieldwork, sorting and analysing material, interpretation and presentation. “Everybody has to go through every different activity to get a full understanding of how we do it and why,” says Douglass.

Yet there’s one area in which team members have yet to be involved: co-authorship and publication. Manahira, for example, has never had his name on a paper, although he would like to. Douglass has published four papers based on the project; although several acknowledge team members and their unique experience, they do not list them as co-authors.

“It can be a bit of a minefield,” says Sarah



Maputo in Mozambique is participating in the climate-resilience research project, FRACTAL.

“THE END PRODUCT OF CO-PRODUCTION DONE WELL IS ALMOST IN THE INTANGIBLE.”

► Sub-Saharan Africa’s urban population is projected to double in the next 25 years. And climate change is expected to hit the continent particularly hard. But although most African governments are aware of climate change, many efforts to inject climate science into city planning have had limited success. Partly, that is because many initiatives assume that the main reason why cities fail to make their development plans climate-proof is a lack of knowledge. If they could just get more accurate information, they could better prepare for what lies ahead. But solving climate-related challenges in developing countries requires more than just climate predictions, says Chris Jack, a climate modeller at the University of Cape Town in South Africa and one of the programme’s lead scientists. “The focus on providing information has distracted from the need to co-produce solutions,” he says.

FRACTAL tries to address this by setting out to understand what cities actually need — and then working with city planners and local scientists to co-produce the missing knowledge.

The learning labs are one of the main ways in which the project explores this shared knowledge. FRACTAL hosts the labs in each city every six months, and has produced a list of burning issues that residents of each area are most concerned about. In Lusaka, these include flooding

and the unregulated use of groundwater, as well as poor sanitation and erratic water supplies. FRACTAL also studies how decisions are made in the cities. This doesn’t happen in the way that most scientists think, says Jack. “We imagine that you bring the data together, you integrate it and you make a decision. In reality, it’s much messier than that.”

One major realization, he says, was that decisions that influence cities’ climate resilience often fall outside the remit of local legislators. Investments in large infrastructure projects in Africa, such as power stations and water pipelines, are sometimes made by development banks and global agencies, meaning that city authorities have limited involvement.

So FRACTAL stakeholders have produced climate-risk narratives for each of the cities, based on different climate-change scenarios described by international models. Many of these deliver different, and sometimes contradictory, predictions. The Lusaka scenarios all involve a warmer city, but the rainfall varies: in one scenario it is drier than now; in another, the rainfall is unchanged; and in a third, there is greater variability in rainfall, with prolonged periods of drought and heavier downpours.

Each scenario describes what will happen to important variables such as water supplies,

flooding and sanitation. They are particularly useful in helping stakeholders to visualize the futures they need to prepare for, says Mununga Mungalu, a senior engineer at Lusaka Water and Sewerage. He says that the narratives have helped to guide his company’s corporate plans and disaster preparedness. “It has changed the culture,” he says.

And the FRACTAL programme has helped people around the region to share knowledge and experience, Mungalu adds. “Most of our institutions plan in silos, and FRACTAL has let us have a voice across these institutions.”

Everyone is on a learning journey, says socio-economic planner Brenda Mwalukanga, FRACTAL’s embedded researcher in Lusaka. All involved accept that they have something to learn from each other, and things are changing, she says. “I have seen scientists engage more with civil society about city governance, and non-science stakeholders request training in climate science.”

The difficulty, then, lies in finding ways to capture and communicate the knowledge being produced by the project. There are different outcomes in each city, involving various groups of stakeholders. “The end product of co-production done well is almost in the intangible,” says Jack. “It’s in the fact that you have connected people and started those discussions.” ■

Cassandra Willyard is a freelance science journalist in Madison, Wisconsin. **Megan Scudellari** is a science journalist in Boston, Massachusetts. **Linda Nordling** is a freelance journalist in Cape Town, South Africa.

JAMES ORTWAY/PANOS

COMMENT

METRICS Patients, farmers and more must co-create tools to evaluate and incentivize **p.32**

INTERNET The rise of Reddit — social software or social malware? **p.34**

MATERIALS The interplay of minerals security and US foreign policy **p.36**



OPEN ACCESS Paywall documentary hits screens as Plan S lands **p.37**

STARWORKS NETWORK



Children with artificial limbs and their carers talk to researchers and industry representatives about improving prosthetics.

Co-production from proposal to paper

Three examples show how public participation in research can be extended at every step of the process to generate useful knowledge.

GARY HICKEY

Share power in five ways

Senior public-involvement manager at INVOLVE, a UK health-research advisory group

A project that is co-produced is one in which researchers, practitioners and the public together share power and

responsibility for the work throughout. The 'whys' of this process are self-evident: patients and the public have the right to be more than just participants in research, and their involvement can lead to better outcomes.

Take, for example, the Child Prosthetics Research Collaboration. This project brought together children and their families with the National Health Service,



CO-PRODUCTION OF RESEARCH

A *Nature* special issue

nature.com/collections/coproduction

industry and academia, and was funded by the UK National Institute for Health Research (NIHR). It led to inventions and optimizations that reflected what children and families need. The experts and academics who develop prosthetics would probably never have heard from families and children how a poor-fitting or unattractive limb can limit a child at home, in the classroom and in the playground.

The 'how' of co-production is less obvious. For the past two-and-a-half years, I have worked with colleagues from the NIHR and beyond to develop guidance ►

► on co-production and to establish an international network for patient and public involvement in health research. It is the main part of my role at INVOLVE, the national advisory group in England's NIHR to foster public involvement in health and social-care research.

Our team held workshops, iterative round-table discussions, consultations and a literature review to characterize co-production. Public members felt that many researchers and practitioners claim their work is co-produced, but still do not respect patient knowledge as equally valuable or put in the effort to ensure that the patient voice has true power.

We identified a handful of principles that define co-production. The crucial one is power sharing — no longer do researchers or practitioners make all the key decisions or take on all the responsibilities.

Sharing power depends on building and maintaining relationships across researchers, practitioners and public members of a research group. That means constant reflection on power differentials, and managing these to build trust. For example, the project team for a study to improve online responses to patient feedback baked cakes and communicated using social media to take people away from the work environment. Everyone could engage in these informal activities, giving people a chance to chat and to reduce anxieties. Space for team-building should be explicitly scheduled into the research cycle. Holding meetings in a neutral setting, such as the local library, and providing opportunities for regular feedback, also helped to build open and trusting relationships between team members.

All relevant perspectives and skills must be included. At the beginning of a research project, the team should consider which knowledge, views, experiences and skills are required, and how to ensure diversity and inclusion. Members should collectively ask: which voices are not around the table?

The knowledge of all team members should be respected and valued. For instance, in a project to update a systematic review of stroke physiotherapy (undertaken by the Nursing, Midwifery and Allied Health Professions Research Unit at Glasgow Caledonian University, UK), the working group of stroke survivors, carers, physiotherapists and educators developed a set of rules. To make it easier for everyone to get their voice heard, no one would be allowed to jump into the group discussion without first raising their hand. To avoid individuals dominating discussions, no one was allowed to speak for more than two minutes at a stretch.

Reciprocity is imperative. Everyone should feel that they get something back from working on a project. For patients, this might be bigger and better social networks, access to training, co-publication and co-presentation, more self-confidence,

a sense of contributing to the greater good, or even payment. For example, a project by Newcastle University set out to develop ways to help young people with neuro-disabilities to participate in leisure activities. Researchers, affected children and artists co-produced an animated film to share the results. The film-makers, 'AniMates', continue to make artwork about research projects, and are now collaborating with other researchers and advisory groups.

It can be difficult for researchers to truly share power when universities are often the main recipients of research grants and academics are ultimately accountable for how the money is spent. New sorts of partnership help. For example, the charity Alcohol Research UK funded its own joint project with the University of Bedfordshire in Luton to explore the experiences of older adults in residential alcohol-rehabilitation services, rather than handing the reins entirely to a grant recipient.

Co-production won't just happen because it is a good thing. The way in which research is currently funded and organized is an obstacle to meeting these principles. Policy-makers, funding bodies, institutions, journals, patient advocates and others need to change their practices and cultures to enable the necessary relationships and facilitate the sharing of power.

TESSA RICHARDS

Get patients to review papers

Senior editor at The BMJ and leader of its patient-partnership initiative

At the clinical journal *The BMJ*, patients and patient advocates have an influential role in our day-to-day decision-making. The journal has long championed partnership with patients in health care and, five years ago, we stepped up our advocacy for it. We were stimulated to do so by mounting concern about wasteful and inequitable Western health systems that fail to serve patients well.

We set up an international panel of patients and patient advocates, and asked them what we needed to do to 'walk the talk' on patient partnership in our editorial processes. I was asked to work with the panel and my editorial colleagues to develop and implement a BMJ patient-partnership strategy.

It has attracted much interest. Every week, we hear from patients, health professionals and policymakers from around the world who share our passion for partnership. They are keen to draw attention to their work and to learn more about what we do.

Incorporating patient and public review alongside our conventional peer-review

processes was the first change we introduced: initially for research papers, then for educational and scholarly comment articles, too. We have an open invitation for people to join our database of patient and public reviewers, and around 700 are registered. Editors invite comments from the reviewers whose lived experience matches as closely as possible to the papers under consideration.

A formal study of the initiative is planned, but informal feedback has been encouraging. Editors report that patient and public reviewers provide valuable perspectives that complement those provided by academic reviewers. These include insights into the wider impact of illness — biological, psychological and social — the 'burden' of treatment, how people self-manage conditions, and whether interventions are practicable. Some patient reviewers have asked authors to modify statements that are not backed by strong evidence, to avoid arousing unjustified hope in the patient community. They also flag inadvertent use of perjorative language, such as 'the patient failed treatment'. Authors have told us that they now think about how their research might be seen through the eyes of patient reviewers.

A survey of our patient and public reviewers found that they greatly appreciate the opportunity to comment on *The BMJ*'s papers and to be involved in our work (S. Schroter *et al.* *BMJ Open* 8, e023357; 2018). Many see it as a way to use their experience of illness to help others. We also learnt that we need to explain editorial processes more clearly and communicate with reviewers more often. Our guidance now addresses reviewers' concerns (for example, explaining that it is OK to decline an invitation to review) and we send out regular newsletters thanking reviewers for their support and updating them on developments.

Patient editors and the continued lively dialogue we have with our patient panel help to implement all aspects of our strategy. Foremost is the requirement that authors submitting a manuscript specify whether and how patients were involved in setting the research question, the design and implementation of the study and its dissemination. Patients and patient advocates also write for us, sit on our editorial board, are members of the advisory committees for our conferences, and are involved in the panels that judge our annual awards.

The changes we have introduced are gradually spreading across the BMJ portfolio of journals, and a few other journals have taken similar steps, such as the obstetrics and gynaecology journal *BJOG*. Patients have responded by drawing up a charter calling for patients to be included in the processes of medical journals (see <https://patientsincluded.org>). We believe that their inclusion will help to improve the quality of health research.



Anne Klein (second from right) is a patient advocate on a clinical-trial panel for her son Everett Schmitt (far right), who has severe combined immunodeficiency.

JEFF SHEEHY

Ask patients what to fund

Board member, California Institute for Regenerative Medicine, USA

In 2004, voters in California allocated US\$3 billion in bonds to create the California Institute for Regenerative Medicine (CIRM), which funds research to produce therapies from stem cells. Unusually, patient advocates such as myself wield formal power at CIRM. Of the 29 board members, 12 slots are designated for patient advocates, including the chair and vice-chair. Board members participate in peer review of all grants, including clinical-stage grant applications. Once formal reviews are in, we vote on the final approval of all grants. A patient advocate is also required on each of the 68 clinical advisory panels that guide late-stage projects, such as a CIRM-funded trial for severe combined immunodeficiency.

I was diagnosed with HIV in 1997. For the past three decades, I've been an activist for the rights of people from sexual and gender minorities (LGBT+), and have even coordinated acts of civil disobedience. I have held legislative office in San Francisco and been the communications director of the AIDS Research Institute at the University of California, San Francisco. When I was appointed to the CIRM board, I had no

interest in merely being a cheerleader or in rubber-stamping decisions that could affect people's lives.

I knew to expect pushback. The legislation that created CIRM gave a voice to patient advocates, but scientists had no experience of having to listen. Many researchers doubted that patient advocates could truly participate in decision-making. Over time, however, relationships between advocates and scientist reviewers developed and scepticism abated.

Time and familiarity were key. The grant-review process at CIRM often lasts for a couple of days, with people being brought together over meetings and meals. We got to know each other through robust debates over different approaches to what research to fund; for example, extremely prevalent diseases versus rare ones neglected by pharmaceutical companies. After 12 years of such gatherings, many of the reviewers have become friends and they listen to, and even welcome, input from me and other patient advocates.

These discussions are not academic to us. Expert scientific reviewers often focus on the high risk of failure. Patient advocates are more willing to champion outlier science. We can make informed decisions to accept high risk if it is balanced by the potential for great reward. For instance, I pay special attention to grant applications that receive highly varied scores from reviewers. Our

"Patient advocates are more willing to champion outlier science."

influence has sometimes meant that a risky grant has been funded over a safer one with higher median scores.

For clinical applications, experts might be less enthusiastic when the best possible outcome is only a partial result — such as a person with a spinal cord injury regaining the use of their arms but not the ability to walk. Yet that partial improvement lets people move themselves in and out of their wheelchair to use a car; to type and text; and to lead an independent life instead of requiring round-the-clock care. I believe that, when it comes to discussions on funding, we bring a clearer understanding of the impact on patients.

Patient advocates can also be more sceptical of strategies that consider human physiology but neglect behaviour. Such strategies include 'kick and kill' HIV therapies, commonly supported among scientists, which eliminate the virus but do nothing to prevent reinfection.

Patient advocacy was pioneered in many ways by AIDS activists, and is now formally referred to in Europe as co-production. People living with conditions, and those who care for them, provide context and counterpoints to unchallenged scientific wisdom.

During its tenure, CIRM has awarded almost 1,000 grants and funded 49 clinical trials, as well as trials of gene therapies that saved the lives of ten people who have different, rare diseases of the immune system. The institute has certainly encountered plenty of controversy, but I think that patient advocates helped it to weather those storms and to steer the best course. ■



Traffic around the Democracy Monument in Bangkok, where city plans aim to improve the quality of life.

Craft metrics to value co-production

To assess whether research is relevant to society, ask the stakeholders, say **Catherine Durose, Liz Richardson and Beth Perry.**

Advocates of co-production encourage collaboration between professional researchers and those affected by that research, to ensure that the resulting science is relevant and useful. Opening up science beyond scientists is essential, particularly where problems are complex, solutions are uncertain and values are salient. For example, patients should have input into research on their conditions, and first-hand experience of local residents should shape research on environmental-health issues.

But what constitutes success on these terms? Without a better understanding of this, it is harder to incentivize co-production in research. A key way to support co-production is reconfiguring that much-derided feature of academic careers: metrics.

Current indicators of research output (such as paper counts or the *h*-index) conceptualize

the value of research narrowly. They are already roundly criticized as poor measures of quality or usefulness. Less appreciated is the fact that these metrics also leave out the societal relevance of research and omit diverse approaches to creating knowledge about social problems.

Peer review also has trouble assessing the value of research that sits at disciplinary boundaries or that addresses complex social challenges. It denies broader social accountability by giving scientists a monopoly on determining what is legitimate knowledge¹. Relying on academic peer review as a means of valuing research



can discourage broader engagement.

This privileges abstract and theoretical research over work that is localized and applied. For example, research on climate-change adaptation, conducted in the global south by researchers embedded in affected communities, can make real differences to people's lives. Yet it is likely to be valued less highly by conventional evaluation than research that is generalized from afar and then published in a high-impact English-language journal.

NOT GOOD ON PAPER

There are now many examples of work co-produced by local partnerships that address health inequalities or environmental and social injustice. Today's 'publish or perish' system in academia vastly undervalues outputs from such projects, which often don't come in the shape of a paper.

Examples challenging this include the feature film *Pili* (2018), a ground-breaking co-production project. The women of Miono in west Tanzania make up the ensemble cast of non-actors, 65% of whom are HIV positive; their real stories provide the basis for the film. It came together as part of a research project on global health, led by political economist Sophie Harman at Queen Mary, University of London, that aimed to give a voice and visibility to unseen women on the periphery of world politics.

Another example of co-production that would be underrated by conventional measures is the Massachusetts Institute of Technology's Fab Lab Network. This open community of scientists, engineers, educators, students and artists of all ages is located across more than 1,000 laboratories in some 100 countries. Fab Labs is, in part, a distributed research lab that aims to democratize access to the tools, education and means for invention, to create opportunities to improve lives.

Consider also the Morris Justice Project. Residents of the Bronx in New York City worked with the City University of New York's Public Science Project to challenge the New York Police Department's 'stop and frisk' policy, which had been rolled out to prevent gun violence. Running since 2011, the project combines research, community participation and action. It showed that people in the Bronx were stopped by police 4,882 times in the first year. More than half of the stops involved physical force, but less than one-tenth resulted in an arrest or summons — and only eight guns were found. The research contributed to a city-wide movement, Communities United for Police Reform, to ensure that debates challenging existing policies were informed by robust and locally informed research. This co-produced work helped to reform legislation and supported several landmark class-action lawsuits.

Another example is the Resource Center for Raza Planning at the University of New

Mexico in Albuquerque. Over its 20-year history, the centre has brought together planning researchers, professionals and traditional communities in New Mexico to influence policy decisions on issues such as economic development, land use, water rights and infrastructure. It exists to ensure that traditional communities are sustainable by co-producing research and making sure this is incorporated into policymaking².

The real-world effects of these examples depend on extending the research community³. Although that still makes many academics uncomfortable, people increasingly acknowledge that local, experiential or applied knowledge can enrich the quality and impact of investigations. The work is more responsive, socially relevant and connected to affected communities.

What is missing are ways to measure success in those dimensions — meaningfully, consistently, rigorously, reproducibly and equitably.

REPORTING STANDARDS

To encourage the practices that broaden research communities, we must make those procedures apparent. Then they can be evaluated and, crucially, rewarded.

Reporting standards could go a long way. The CONSORT Guidelines for reporting the results of clinical trials were proposed in 1996, and have now been taken up almost universally, enforced by journals and government funders. Before their adoption, reports of clinical trials were hard to appraise. Similar efforts around co-production would be advantageous.

It is still early days. We cannot assume that we are all on the same page about what it means to co-produce research, especially across different scientific disciplines. Reporting standards around the research process could clarify what is involved when different groups talk about co-production (see go.nature.com/2nzn7xw). They would show how research was planned, conducted and applied⁴.

An emerging strategy is to clearly state the intentions of co-produced work, and evaluate it on the basis of the intentions. If the intention is instrumental — to characterize lay knowledge of local conditions, say — then the metric would be based on the inclusion of that lay knowledge. If the intention is to honour inclusion — encapsulated in the disability-rights call, ‘nothing about us without us’ — then more-appropriate metrics might centre on how participants perceive the quality of their involvement in the work.

Reporting standards should capture the stage of the research process at which co-production occurs⁵. Were the initial research questions defined co-productively? Or did co-production happen later, such as during analysis, interpretation and dissemination of the findings? For example, *The BMJ*

BEST PRACTICE

Two co-production tips

Survey your options. Different groups of extended peers will need to hammer out their own criteria for co-producing research, but examples of good practice and templates to describe intentions and processes will help. Recommendations should be aligned with guidelines on responsible metrics¹⁰. There is precedent. In 2011, the UK Arts and Humanities Research Council’s Connected Communities programme commissioned a group of Durham University academics and community partners to examine and make recommendations on the ethical challenges raised in community-based participatory research (see go.nature.com/2qyh21j).

Support long-term partnerships.

Institutions and funders must put resources into extended peer communities. For instance, the UK Economic and Social Research Council has invested in our Jam and Justice project (www.jamandjustice-rjc.org). This explores how an extended peer community can govern research around positive urban transformations. Similarly, the University of Illinois at Chicago employs community-development workers in its Office of Community and Public Health Practice; they sustain relationships with local organizations to enable community-based research. **C.D., L.R. & B.P.**

now requires that all its journal articles acknowledge whether and how patients or carers were involved in research — a demand that came about through consultation with those communities (see T. Richards, page 30).

TOOLS NEEDED

The extended peer community should play a part in determining any evaluation system. The goal is not to be prescriptive, but rather to clarify the intentions and processes of scientists and other co-producers of research. An accepted suite of criteria helps to document these choices and leads to context-appropriate evaluation (see ‘Best practice’)⁶.

There are only a handful of examples to build on⁷. Co-production tends to function at a small, experimental scale, and generally does not attempt to draw out working principles that other programmes might learn from.

One notable exception is the organization Mistra Urban Futures (for which B.P. serves as the UK lead). It has developed workshops that support peer learning for people working

on co-produced research, a transdisciplinary research school and a handbook, alongside an evaluation methodology for co-production that considers both the quality of the processes and the outcomes achieved⁸. This international centre focuses on how cities and settled areas can grow sustainably, and is led by a consortium of local authorities and academics in Sweden, with partners in South Africa, Kenya and the United Kingdom⁹.

Mistra’s criteria for high-quality co-production include relevance, credibility and legitimacy. Outcomes are categorized in several ways: as effects that can be directly attributed to a programme and as potential broader effects and influences.

The use of proxies to measure outcomes is crucial, yet is underdeveloped. Proxies for social values (such as commitment and a feeling of belonging) can include contributions in kind, time donated to projects and the depth and breadth of resulting personal, inter-relational and system-wide networks.

Another organization of note is Canada’s International Development Research Centre in Ottawa, which funds work aimed at tackling social and health problems in the global south. It has developed a tool for assessing the projects that it supports, which incorporates the views of stakeholders, users and non-scientific beneficiaries in communities.

Co-production doesn’t devalue science, it re-evaluates other ways of knowing. If we want to see more co-production, we need to revise the dominant metrics accordingly. In essence, metrics to assess co-production must themselves be co-produced. ■

Catherine Durose is a reader in policy sciences at the University of Birmingham, UK. **Liz Richardson** is a reader in politics at the University of Manchester, UK. **Beth Perry** is a professor in urban studies at the University of Sheffield, UK. e-mail: c.durose@bham.ac.uk

1. Walker, D. *Public Money Mgmt* **30**, 204–206 (2010).
2. Durose, C. & Richardson, L. *Designing Public Policy for Co-Production: Theory, Practice and Change* (Policy Press, 2016).
3. Funtowicz, S. O. & Ravetz, J. R. *Futures* **25**, 739–755 (1993).
4. Molas-Gallart, J. *Arts Humanit. Higher Educ.* **14**, 111–126 (2014).
5. Richardson, L. in *Handbook of Social Policy Evaluation* (ed. Greve, B.) 119–138 (Elgar, 2017).
6. Pain, R. et al. *Mapping Alternative Impact* (N8 Research Partnership, Durham Univ. & Economic and Social Research Council, 2015).
7. May, T. & Perry, B. *Cities and the Knowledge Economy: Promise, Politics and Possibilities* (Routledge, 2018).
8. Polk, M. in *Disciplining Interdisciplinarity* (ed. Bammer, G.) Ch. 51 (ANU Press, 2013).
9. Perry, B., Patel, Z., Norén Bretzer, Y. & Polk, M. *Polit. Gov.* **6**, 189–198 (2018).
10. Wildson, J. et al. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management* (2015).

Competing financial interests declared; see go.nature.com/2rjy8gp for details.



Alexis Ohanian (left) and Steve Huffman founded the social-media website Reddit in 2005.

INTERNET

The rancorous rise of Reddit

Timo Hannay extols a history of the website's evolution in our tumultuous era.

According to Internet-analytics company Alexa.com, the websites with most traffic from the United States are Google, YouTube, Facebook and Amazon (which owns Alexa). The rest of the top ten is composed of other familiar names — Yahoo!, Twitter, Wikipedia, Instagram and LinkedIn. But one entry might surprise. Not only is it listed at an impressive number five, but it beats all the others hands down in terms of time spent by each user. Welcome to Reddit.

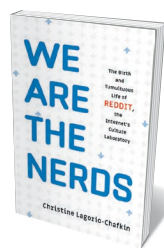
We Are the Nerds by journalist Christine Lagorio-Chafkin traces Reddit's emergence in 2005, and its evolution as a website, company and social phenomenon. Ostensibly, it's the story of co-founders Steve Huffman (the technical brains) and Alexis Ohanian (the showman). But it is really three tales in one.

The first story is that of a scrappy start-up destined for web domination. Superficially, this resembles the legend of Apple founder Steve Jobs: foundation, separation, return and redemption. Reddit's sale to New York-based magazine publisher Condé Nast in 2006, just over a year after launch, bestowed wealth and credibility on its young founders, but it was a cultural mismatch. Huffman and Ohanian lost heart; by 2010, both had left, their friendship strained. Feted replacements such as

investor Ellen Pao never quite embodied the Reddit spirit. Crises followed, from staff rebellions to a spate of revenge porn between users. Neatly for the narrative arc, a 2015 rapprochement led to Huffman and Ohanian's surprise return — and the revitalization of their wayward creation.

The second story concerns the early-twenty-first-century technology industry. Here, Reddit is a node in a network of technologists, entrepreneurs and iconoclasts seeking to reshape the world. The reader feels like Forrest Gump, stumbling from one remarkable event or person to the next.

Science publishing makes an appearance, albeit a tragic one. Hired at the start of Reddit's journey, programmer Aaron Swartz quickly became more taken with campaigning than coding. Incensed by publishers' paywalls, he covertly downloaded millions of academic articles, and was caught. The ensuing legal



We Are the Nerds:
The Birth and
Tumultuous Life
of Reddit, the
Internet's Culture
Laboratory
CHRISTINE LAGORIO-
CHAFKIN
Hachette (2018)

battle ended in 2013, when this principled, sensitive young man killed himself, aged 26.

As this second narrative unfolds, readers might lose track of the vast array of walk-on parts. But for anyone familiar with the names, it's a who's who of nerd aristocracy, from Paul Graham, co-founder of Y Combinator in Mountain View, California (the start-up incubator that begat Reddit) to Chris Anderson, former editor-in-chief of technology magazine *Wired*. It is also a reminder of how few people comprise the circles of influence in the parochial but powerful world of the web.

The third story traces the rise of social media from the perspective of one of its most important players. If Facebook lets users cultivate online personas and Twitter enables them to broadcast random thoughts to the world, Reddit was built to foster discussion. Whatever one thinks about the social costs and benefits of such services, they are no longer mere geeky distractions. They are central to the perceptions of billions, and have become cultural and political battlegrounds.

In August 2012, it all seemed positively wholesome. At the University of Virginia in Charlottesville (where Huffman and Ohanian met), then-US president Barack Obama took part in an Ask Me Anything or AMA, a Reddit staple in which anyone from A-listers to the

REDDIT

terminally obscure (I've done two) answers questions. Obama was a natural, typing his own answers and signing off with a Reddit catchphrase: "NOT BAD!". The crowd went wild. Four years later, it was all very different.

Reddit, as *We Are the Nerds* shows, was always a venue for the edgy and degenerate, fostered in part by its anonymity. But by 2016, some of this was going mainstream. The forum (or 'subreddit') r/The_Donald had become an important cheerleader for a divisive US presidential campaign. The volunteer moderators kept just inside the rules. It became a prolific disseminator of misleading memes — with consequences that everyone now knows but no one yet fully comprehends. If Obama was the presidential incarnation of change-the-world techno-optimism, Trump now personified a revenge of the trolls.

In August 2017, white supremacists and opposition demonstrators went head-to-head in Charlottesville. Huffman was furious, and the incident triggered a clampdown on certain far-right groups across Reddit. This was a major milestone on the journey from the site's freewheeling origins to a dawning realization that online communities, like societies, need rules. An obvious question is why unaccountable individuals such as Huffman (or Facebook founder Mark Zuckerberg, or Twitter's Jack Dorsey) should be the ones setting them.

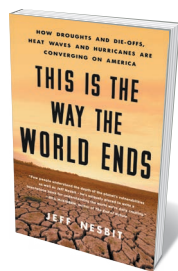
The story of social media is ironic. The most powerful decentralizing technologies in history — the Internet and the web — have led to the greatest concentrations of power. Friction-free information and the death of distance have not ushered in a new Enlightenment, but enabled every crackpot belief and bile-drenched enmity to gain adherents. Technologists, anxious to avoid any 'single point of failure' in their systems (the reason everything from disk drives to data centres is duplicated) have built single points of failure for society. A well-aimed post or algorithmic tweak can mislead, enrage and divide on a national or global scale. At its all-too-common worst, this is not so much social software as social malware.

The main story of the book ends on a high. Huffman is the boss of a major website valued at well over US\$1 billion. Ohanian, Reddit's first promoter and now its executive chair, is a celebrity (and married to tennis phenomenon Serena Williams). To paraphrase Jobs, both have helped to put a dent in the Universe.

But this is no happily-ever-after fairy tale. *We Are the Nerds* describes how Reddit began. The real story is how the site and its ilk will change the world. On that, we're still in Act One — and the story is being written by us all, one thoughtful blogpost or belligerent tweet at a time. ■

Timo Hannay is the founder of education technology company SchoolDash.com, based in London.
e-mail: timo@hannay.net

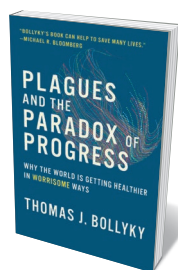
Books in brief



This Is the Way the World Ends

Jeff Nesbit THOMAS DUNNE (2018)

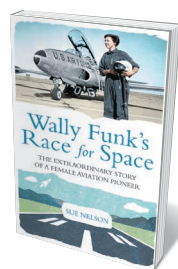
Environmental expert Jeff Nesbit delivers a scientifically rich overview of how the impacts of climate change are affecting natural resources in the here and now. He reveals how oceanic and atmospheric shifts are triggering losses in species from pollinating insects to phytoplankton, fatal heatwaves are becoming regional norms and water stress could spark new waves of mass migration. Nesbit's blueprint for surviving these systemic issues — centring on efficient resource use, innovation and infrastructure — is arguably sketchy, but overall this is a cogent analysis of a creeping crisis.



Plagues and the Paradox of Progress

Thomas J. Bollyky MIT PRESS (2018)

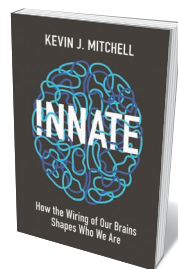
From polio to tuberculosis, infectious diseases are no longer the leading cause of death in any region. Yet this triumph is paradoxical, argues Thomas Bollyky in this rich, incisive study. Bollyky, director of the Global Health Program at US think tank the Council on Foreign Relations, shows that in too many lower-income countries, any gains in public health are counterbalanced by poor health-care systems, illiberal governance, low employment, rampant urbanization and booming populations. A thoughtful reminder of the social, economic and political complexities inherent in sustainable public health.



Wally Funk's Race for Space

Sue Nelson WESTBOURNE (2018)

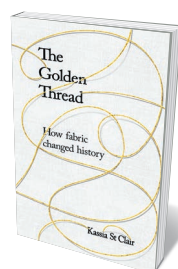
In 1961, as NASA made superstars of astronauts such as John Glenn, their medical supervisor, William Randolph Lovelace II, was secretly training 13 female flying aces for space. The 'Mercury 13' programme was axed, but the ambitions of trainee Wally Funk never died. In this compelling portrait, space journalist Sue Nelson reveals how Funk (now nearly 80) became the first female US aviation-safety inspector, has hobnobbed with luminaries such as Italian astronaut Samantha Cristoforetti, and is in training for the first Virgin Galactic flight. As Nelson notes: "What a life she has had while fighting to right a wrong."



Innate

Kevin J. Mitchell PRINCETON UNIVERSITY PRESS (2018)

The nexus of neuroscience and genetics can be murky. Not so in neuroscientist Kevin Mitchell's study on human diversity, which probes with clarity and balance how variation in our genetic program causes variation in outcome. Mitchell reveals that environmental effects tend to amplify, not counteract, innate differences. He uses that framework to examine psychological domains such as perception, conditions including schizophrenia, and the dubious ethical and social implications of 'designer babies' and other trends. A powerful antidote to genetic determinism.



The Golden Thread: How Fabric Changed History

Kassia St Clair JOHN MURRAY (2018)

Fabrics are knitted into human history, from the Silk Road to the mechanical looms of the Industrial Revolution. Here, design writer Kassia St Clair explores the connection in 13 beautifully wrought stories. We visit a cave in Georgia's Caucasus Mountains where dyed fibres more than 30,000 years old have been discovered; goggle at the starched intricacy of sixteenth-century lace ruffs; flinch over astronauts' nappies and the sodden sleeping bags of early polar expeditions; and savour the idea of materials spun from spiders' webs. A joyful commingling of text and textiles. **Barbara Kiser**

Minerals and manifest destiny

Is the US Department of the Interior imperialist? **K. John Holmes** investigates.

Despite its name, its remit is international: the US Department of the Interior (DOI) is not a body solely devoted to managing land and natural resources within US borders. Historian Megan Black's *The Global Interior* would have us believe that that has been disastrous. Despite scandals and disasters stretching back to at least the 1920s, the DOI retains a fairly innocuous reputation. Black argues that it has used that characterization to satisfy insatiable US demand for minerals from copper and tin to bauxite and lithium, and to enable the expansion of US imperialism.

The Global Interior tracks the scope of the DOI's minerals legacy from the arid US West to Alaska and island territories, South America, the Middle East, and eventually to the ocean floor and outer space. Black ends back in the US West, where ongoing struggles between Native American interests and the mining and energy-extraction industries close the circle. She reveals a complex strategy, ranging from securing energy resources and industrial ores for war efforts to aiding US and international companies through resource assessments, diplomatic activities and direct aid. The DOI has, for instance, assisted operations by the now-defunct Bethlehem Steel in Cuba, and provided technical aid for mica mining in Brazil.

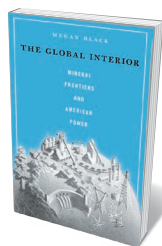
The book lays out the international scope of the US government's resource-directed activities, although some environmental historians might be surprised by the oversized role Black gives to the DOI. Even less recognizable to some will be Black's thesis that a country can expand its domain through environmental stewardship. Ultimately, although provocative, her narrative jars with the rich history of the DOI's science, analysis and resource assessments over its nearly 170 years, in my view. As when analysing any piece of thought-provoking scholarship, disagreeing is as much of the journey as agreeing.

Black does, throughout, demonstrate a keen sense of the uneven balance of power between the US government and peoples living on resource-rich lands, and lays out how, in pursuit of minerals, the United States has exploited and marginalized the capabilities and interests of these people around the world. She also describes how the DOI, in its earliest days during US westward expansion, offered settlers lands expropriated from indigenous people. Even nineteenth-century explorer and scientist John Wesley Powell — whom I have promoted for foreshadowing today's climate-assessment

practices — advocated removing Native Americans from their lands in the arid West at the same time as documenting and celebrating their cultures.

However, Black reduces the DOI and its history to a resource and development branch of the federal government. In doing so, she ignores the vast reach of its domestic and international scientific and non-minerals activities since the beginning of the twentieth century. Its domestic remit, she contends, was achieved at the close of the nineteenth century, when several of its sub-agencies collectively disposed of public lands, and contained Native American peoples on reservations.

Yet at that time, the department had already launched a massive domestic scientific programme led by its US Geological Survey (USGS), producing impartial, publicly available monitoring and analysis of the nation's lands, subsurface resources, natural hazards and water quantity and quality. From the 1880s, the USGS began developing the National Map — a topographical chart of the



The Global Interior: Mineral Frontiers and American Power
MEGAN BLACK
Harvard University Press (2018)

48 contiguous states, the first draft of which was finally completed in 1991. It gauged thousands of rivers and streams, and produced more than 100,000 publications on geology, biology and ecosystems, coasts and oceans, energy, minerals, natural hazards and water.

The DOI's Fish and Wildlife Service, whose precursor was established in 1871, has a mission ranging from enforcing federal wildlife laws to fisheries management. The National Park Service manages more than 400 sites and wilderness reserves (see E. Carr *Nature* 535, 34–36; 2016). The Bureau of Reclamation operates hundreds of irrigation and hydropower dams in the US West.

Black's view of the DOI's twentieth century is very different. Using politically charged language, she contends that unnamed US leaders reoriented the department to focus on “ever-widening horizons, including formal imperialism” and spread its operatives to develop a “mineral intelligence base” that would expand the country's dominance. She reduces Second World War efforts to secure strategic materials to a lust for “materials needed for armament”. At the same time, she presents little systematic analysis to support her contention that the DOI's focus was actually on globalization and the pursuit of empire-building.

A key element of DOI history, which Black does bring forward, is the difficulties that inevitably arise for an agency that attempts

BETTMANN/GETTY



Bethlehem Steel Mill in Johnstown, Pennsylvania, in 1937. The company used iron from Cuba.

to manage, regulate and promote the use of all of a nation's natural resources. The 2010 Deepwater Horizon oil spill, for instance, resulted in the dissolution of the DOI's Minerals Management Service; it became clear that one service could not oversee offshore oil and gas development, collect royalties and enforce regulations and safety. As Black notes, the incompatibility of the DOI's multiple missions — for instance, those concerning resource development and environmental protection — led to the foundation of competing federal departments and agencies, such as the Forest Service, Environmental Protection Agency and Department of Energy. In my view, this is less a failing of the DOI than a natural evolution: the emergence of spin-off agencies in response to perceived need represents the democratization of science.

I reflect on this book following the mid-September Global Climate Action Summit in San Francisco, California, organized by outgoing state governor Jerry Brown. Brown was nicknamed 'Governor Moonbeam' during his first tenure in the post more than 40 years ago, in part for his embrace of Earth-observing satellite technologies. Landsat — the Earth-observation programme that emerged from a joint enterprise of NASA and the USGS — provided a scientific base from which to improve understanding of resources and the environment. However, in Black's telling, it has been "a tool to further capitalist exploitation", embraced by an "array of well-meaning scientists and unscrupulous dictators".

Given California's economic reliance on the technology industry, Brown's advocacy of high-tech monitoring in pursuit of an aggressive environmental agenda might look self-serving. His vision might one day even be called an expansion of the Californian empire. And it is true that environmentalism should never be immune to critiques of its potential to suppress poorer countries' pursuit of development and opportunity. But to view the development of US capabilities in science and technology over the DOI's long and complicated history solely through the lens of expansionism, greed and imperial tendencies belies the complexities of the world we all live in and the fundamental part that scientific progress plays. ■

K. John Holmes is the director of the *Board on Energy and Environmental Systems at the National Academies of Sciences, Engineering, and Medicine in Washington DC*. His interests are in energy and the environment, both modern and historical.
e-mail: johnes@nas.edu



Helena Asomoah-Hassan, university librarian at KNUST, Ghana, being interviewed for *Paywall*.

PUBLISHING

Open access — the movie

Richard Poynder views a documentary on the tug of war over paywalls in scholarly publishing.

Billed as a documentary, *Paywall* would be more accurately described as an advocacy film. Its intention seems to be to persuade viewers that the paywalls that restrict access to journal content online are an unnecessary hang-over from the print era, and now serve only to perpetuate the excessive profits that legacy publishers such as Elsevier, Wiley and Springer Nature make from the public purse.

The film makes a convincing case that the paywall system creates problems — and that universal open access (OA) to scholarly articles would be better for society. But it fails to adequately explore the thorny challenges that arise with OA publishing. These include the fact that the publishers castigated would continue to dominate scholarly communication in an OA world; the increasingly expensive 'pay-to-publish' model, which substitutes inequities in access for inequities in affording publication; and the rise of predatory publishing. And although *Paywall* acknowledges that current reward systems have slowed the progress of OA publishing, it does not

Paywall: The Business of Scholarship
DIRECTOR: JASON SCHMITT
Open Society Foundations (2018)

address the puzzling question of why academics have proved so reluctant to make copies of their published papers freely available in their

institutional repositories.

Paywall features more than 70 interviews. People represented include: Richard Wilder, associate general counsel at the Bill & Melinda Gates Foundation; Wikipedia Library head Jake Orlowitz; and Alexandra Elbakyan, founder of Sci-Hub (a website that offers free access to more than 70 million illegally downloaded academic papers). Rachel Burley, publishing director for BioMed Central and SpringerOpen, speaks for Springer Nature.

The film ranges over issues such as journal price inflation, researcher evaluation and impact factors, and the disparity of access between the predominantly wealthy global north and the mostly lower-income global south. The film is funded by the Open Society Foundations in New York City, which was created by ▶

► philanthropist George Soros in 1993, and was instrumental in the formation of the OA movement.

Director Jason Schmitt — a scholar of communications and media at Clarkson University in Potsdam, New York — made the film to bring the discussion to the public at large. Yet most of the screenings are scheduled at universities, so how broad an audience it will find is an open question.

Schmitt wrote to me: “Publishing top-tier research journals is complex and costly. I know publishers provide an important service. But I feel that at the current technological bandwidth, we don’t need the sheer number of journals controlled by large publishers.” He describes the scholarly publishing market as a US\$25.2-billion-a-year industry. Heather Joseph, executive director of the global OA advocacy group the Scholarly Publishing and Academic Resources Coalition, puts the figure at \$10 billion.

The film singles out Elsevier for most criticism, eliding the fact that the company is simply more successful than most for-profit legacy publishers at doing what they all do. Schmitt wrote me that he tried to achieve balance, but that Elsevier declined to be part of the film, so it was unable to “show the positives and attributes of their business model”. Instead, the witness for the defence is Will Schweitzer, product-development director at the American Association for the Advancement of Science in Washington DC, publishers of *Science* and other journals. He says: “Do we act effectively as a responsible midwife for these important scholarly concepts or ideas, and make them accessible to the world and distribute them, and reinvest in the community? I would say yes.”

Subscriptions, Schmitt argues, unnecessarily restrict access to research. Moreover, prices routinely increase faster than inflation — and library budgets — so journal subscriptions are regularly cancelled, and paywalls grow.

Paywalls hit researchers from the global south hardest. A 2001 World Health Organization (WHO) survey found that 56% of research institutions in very low-income countries had no subscriptions to international scientific journals. To address this, global agencies worked with major publishers to offer researchers in poorer countries free or low-cost access to articles. Initiatives include the Hinari Access to Research for Health Programme, run by the WHO, and Access to Global Online Research in Agriculture, run by the Food and Agriculture Organization of the United Nations (these programmes and others have now been subsumed under Research4Life). Yet these initiatives are regularly criticized for creating dependency and “commodifying legitimacy”.



JIM RICHARDSON/NGC/GETTY

The University of Oxford, UK, ran out of funds for some open-access publication charges early this year.

The film offers telling examples. Nigerian physician Ahmed Ogunlaja, for instance, explains that local doctors are constantly confronted with paywalls. Another interviewee — Tom Callaway, head of outreach to universities at open-source software company Red Hat in Raleigh, North Carolina — relates that he could not afford to research his wife’s pulmonary embolism. Without a subscription, each paper costs an average of \$30–40, and it is not possible to know whether they are relevant before paying.

I agree with the film that universal OA is far preferable to subscriptions. Combined with open data, it would make science more efficient, not least because more scholars, independent researchers and citizen scientists would be able to contribute to and build on published work. Greater openness could also help to address problems of reproducibility, fraud and research misconduct. And the increasingly interdisciplinary work necessary to address grand societal challenges — from climate change to food security — is better enabled by OA.

The film mentions ‘green’ OA (in which researchers deposit copies of their own papers in online repositories), but seems more focused on ‘gold’ OA, in which publishers make papers freely available.

The weakness of *Paywall* is that it fails to adequately address the challenges of OA. Among the biggest are article-processing

charges (APCs). The now-dominant OA model pioneered by publishers PLOS and BioMed Central, both founded in 2000, demands that authors or their funders pay APCs to make work freely available. But many cannot afford the charges, even at leading universities in wealthy nations. Legacy publishers all now also offer gold options that set APCs at levels designed to preserve current profits. Thus, the very publishers that *Paywall* criticizes will continue to dominate, because (as the film points out) researchers have incentives to publish in their prestigious journals. And for those in the global south, APCs are invariably unaffordable. Waivers are sometimes available, but authors often find they are not eligible. The problems of both affordability and equity will persist.

The film also fails to discuss other pressing issues. These include a lack of consensus on exactly what OA is and how it should be achieved, and the continuing indifference to it in the research community — consider that many academics do not self-archive their papers even when mandated to do so. Moreover, because many OA papers have no licence attached, they are susceptible to being placed behind a paywall later, making openness a fragile condition. It’s ironic, too, that the most successful OA initiative is Sci-Hub.

As a piece of advocacy, *Paywall* is compelling enough to attract new converts. It will not, however, educate the public in the complexities of open access. ■

Richard Poynder is an independent blogger at Open & Shut?
e-mail: richard.poynder@cantab.net

Correspondence

Genetic testing can aid pet breeding

Genomics testing offers benefits for pets beyond health improvements (see L. Moses *et al. Nature* 559, 470–472; 2018). It can help breeders to maintain genetic diversity in populations, for example, by probing ancestry or exposing undesirable recessive traits.

The non-profit initiative Harmonization of Genetic Testing for Dogs, developed by the International Partnership for Dogs, now lists around 70 commercial test providers worldwide on its portal. There, users can find details about companies' quality measures and genetic counselling services, for instance (see go.nature.com/2xsqgnd).

Breeding organizations run genomic-selection programmes for a variety of complex traits in livestock. For example, it is possible to select for lower methane emissions and for improved adaptation capabilities (B. J. Hayes *et al. Trends Genet.* 29, 206–214; 2013).

With careful attention to appropriate markers, phenotype characterization, statistical computation and definition of reference populations, genetic testing can guide breeding decisions and strengthen populations (see go.nature.com/2lweoak).

Gregoire Leroy AgroParisTech, Paris, France.

Brenda N. Bonnett International Partnership for Dogs, Sollentuna, Sweden.

Katariina Mäki Finnish Kennel Club, Espoo, Finland.
gregoire.leroy@agroparistech.fr

Biodiversity: broaden valuation

It is a shame that debates on biodiversity policy are much narrower in some countries than those fuelled by the wide range of voices in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem

Services (IPBES; see *Nature* 560, 423–425; 2018). The United Kingdom, for instance, retains the model of natural science in alliance with economists who specialize in the monetary valuation of 'pieces of natural capital'. The nation must learn from IPBES if it is to address crucial aspects of biodiversity loss.

Economics can contribute much more than techniques and studies of valuation. Changes in land use, the way in which food and energy are produced, and consumer demand for different types of product, for example, can all cause biodiversity loss.

Putting a price on 'natural capital' speaks more to policymakers than to most people's reasons for valuing nature. We need to find a better way to mobilize support for biodiversity conservation — for example, by defining concepts such as services and diversity in terms that are more meaningful to the public.

Many participants in the research network that we coordinate, Debating Nature's Value (see go.nature.com/natval), would like to see UK researchers and policymakers adopt the IPBES approach.

Victor Anderson*, **Aled Jones** Anglia Ruskin University, Cambridge, UK.

Rupert Read University of East Anglia, Norwich, UK.

**Declares competing interests (see go.nature.com/2rbtcwm for details).*

victor.anderson@anglia.ac.uk

Biodiversity: how old is each coinage?

The debate around which framework to use to value biodiversity (see *Nature* 560, 423–425, 2018) could stem from the relatively recent coining and adoption of the concept of nature's contribution to people (NCP; S. Diaz *et al. Science* 359, 270–272; 2018).

Google Scholar returns only 19 hits for NCP and

nearly 100,000 for ecosystem services, mainly because the latter has been in use for much longer. By contrast, this year's summary for policymakers in the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) Europe and Central Asia assessment received 115 and 37 hits, respectively. Given that assessments by IPBES synthesize and build on large bodies of existing scientific and other types of information, the discrepancy in the numbers could imply that the two concepts are used interchangeably.

It takes many years of careful work, peer review and weighing of evidence for a conceptual framework to become widely adopted. The term ecosystem services had its breakthrough at the time of the Millennium Ecosystem Assessment, after some 15 years of development. Much effort has since been spent working with governments to mainstream the concept to underpin action.

It will be difficult for academics and governments to adopt a new paradigm without a proper, rigorous test of its utility. We are convinced that the ecosystem services and NCP world views can be reconciled, and ultimately both need to be endorsed. The degradation of ecosystems and loss of biodiversity remain pressing problems however they are conceptualized.

Jim Harris Cranfield University, Bedfordshire, UK.

Janne S. Kotiaho University of Jyväskylä, Finland.
j.a.harris@cranfield.ac.uk

Overhauling China's electricity sector

China is experimenting with electricity markets to hasten its transition to a clean-energy system. Currently, its annual generation plan allocates roughly the same operational hours to coal power plants irrespective of their cost or

efficiency. Eventually, this will be supplanted by wholesale markets, allowing more-flexible operation of its power system within and across provinces and regions. This is an important step.

The evolution of electricity markets in China will not be swift, painless or linear, thanks to political and economic obstacles as well as entrenched stakeholder interests. But if designed and governed well, markets hold promise for resolving the political challenges that will otherwise frustrate China's transition to a clean-energy system. They can accelerate the replacement of relatively cheap coal-fired energy with renewable energy as the increasing scale of solar and wind installations drives their costs to parity.

Other countries' experiences are useful for reference. But China's electricity markets will ultimately need to develop along a trajectory that is adapted to the nation's particular conditions and challenges. These include long distances between electricity resources and demand centres, the continued need for large-scale investment, the political need for an orderly transition, and China's distinctive approach to governance and regulation.

Markets that allow electricity to be traded and coordinated across regions that cover long distances are likely to be more feasible in China than they have been in the United States or Europe.

Jiang Lin University of California, Berkeley, USA.
lin.jiang@berkeley.edu

CONTRIBUTIONS

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at <http://go.nature.com/cmchno>.

Blood flow forces liver growth

Increases in biomechanical forces in the liver's blood vessels have now been shown to activate two mechanosensitive proteins. The proteins trigger blood-vessel cells to deploy regenerative factors that drive liver growth. SEE LETTER P.128

SINA Y. RABBANY & SHAHIN RAFII

The molecular pathways that initiate and sustain liver growth during development and after injury are orchestrated in part by a balanced supply of stimulatory and inhibitory factors secreted from specialized liver sinusoidal endothelial cells (LSECs), which line the organ's blood vessels^{1–4}. But it is unclear how the liver vasculature senses the need to produce these endothelial-cell-derived (angiocrine) growth factors, such as hepatocyte growth factor (HGF) and Wnt proteins, to guide proper organ growth⁴. On page 128, Lorenz *et al.*⁵ show how mechanical forces created by the passage of blood through the liver activate signalling pathways that promote the production of angiocrine factors and the proliferation of the organ's main cell type, hepatocytes, in mice.

Mechanosensing in the liver depends on the amount of blood delivered by the portal vein and the hepatic artery, and on the tensile strength of blood-vessel walls, which is imparted by collagen fibres. The net blood flow (perfusion) subjects LSECs to two major forces⁶. First, mechanical distortion and tension of the vessel wall owing to blood pressure results in cyclic stretch in the cells. Second, friction arising from viscous blood flow over the vessel wall causes fluid shear stress. These synergistic biomechanical forces lead to upregulation of various mechanosensing proteins, inducing LSECs to produce angiocrine factors such as nitric oxide and reactive oxygen species that act to modulate the vasculature, together with 'stronger' angiocrine factors that stimulate hepatic regeneration. However, the mechanism(s) by which biomechanical forces activate the strong angiocrine function of LSECs to choreograph hepatic proliferation have not been elucidated⁷.

Lorenz *et al.* set out to investigate these mechanisms using mouse embryos removed from mothers and cultured *ex vivo*. They first observed that an increase in the liver's growth rate over different developmental stages correlated with enhanced blood perfusion through the organ. Most proliferating hepatic cells in the developing liver were localized to regions that had been perfused, and the researchers found that the level of vascular perfusion correlated with the level of activation of two receptor

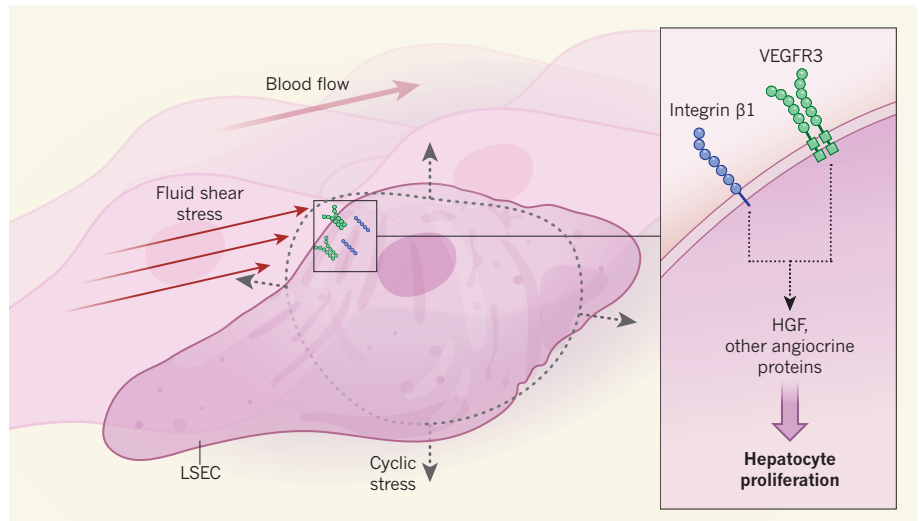


Figure 1 | Biomechanical forces mediate proliferation of liver cells. Blood flow exposes liver sinusoidal endothelial cells (LSECs), which line blood vessels in the liver, to two types of force — fluid shear stress, caused by friction across the cells, and cyclic stretch, caused by dilation of LSECs as the vessels expand. Lorenz *et al.*⁵ report that these forces activate two mechanosensing receptor proteins on LSECs: integrin $\beta 1$ and VEGFR3. Through as-yet-unknown mechanisms, activation of these proteins promotes expression of the gene encoding hepatocyte growth factor (HGF), along with other LSEC-derived (angiocrine) proteins. HGF is secreted by LSECs, and promotes proliferation of the liver's main cell type, hepatocytes.

proteins on LSECs that sense and respond to force — integrin $\beta 1$ and vascular endothelial growth factor receptor 3 (VEGFR3). In turn, these proteins promoted secretion of the key angiocrine factor HGF (Fig. 1).

The authors then modified perfusion rates in cultured embryos by using drugs to halt or increase the fetal heartbeat. Blocking liver perfusion reduced HGF secretion and resulted in diminished hepatic growth — as did deleting the genes that encode integrin $\beta 1$ or VEGFR3 in LSECs in embryos *in vivo*. By contrast, enhancing the rate of blood perfusion induced the secretion of HGF, and this was again mediated by integrin $\beta 1$ and VEGFR3.

Lorenz *et al.* then turned to livers removed from adult mice and cultured *ex vivo*. They increased perfusion in the livers by injecting a buffer solution or by removing 70% of the liver, which redirects a large volume of fluid to the organ's remaining lobes at high pressure. They measured the perfusion rate using an imaging technique called contrast-enhanced ultrasonography. Enhanced perfusion led to increased LSEC diameter, increased blood

volume and flow and so increased shear stress, leading to higher activation of integrin $\beta 1$ and VEGFR3.

Together, these experiments provide evidence that, in mice, activation of mechanosensors in blood vessels in both the fetal and adult liver triggers angiocrine signalling to promote hepatocyte proliferation — presumably, to enable liver growth during embryonic development and maintenance and regeneration in adults. Next, Lorenz *et al.* turned to human cells cultured *in vitro*. Here, too, mechanical stretching of LSEC-like cells or antibody-dependent activation of integrin $\beta 1$ led to a robust increase in secretion not only of HGF, but also of other angiocrine factors. These secreted factors promoted the proliferation and survival of human hepatocytes grown *in vitro*. Finally, the authors showed that, in metabolically healthy people, increases in systemic blood pressure correlated with significantly larger livers.

Lorenz and colleagues have used sophisticated approaches to link mechanical forces to the induction of angiocrine-mediated liver development and growth. However, several

issues remain unresolved. For instance, the cyclic stretch that LSECs undergo *in vivo* when the vessel widens after exposure to accelerated perfusion is biaxial — that is, the cell is stretched both along the direction of the vessel and sideways. By contrast, Lorenz and colleagues' *ex vivo* and *in vitro* experiments imparted only uniaxial cyclic stretch⁸. This difference might bias the signalling and angiocrine outputs the group observed. Whether other vascular mechanosensor receptors have a role in the induction of angiocrine factors also needs to be elucidated⁹.

In addition, the role of this biomechanically responsive pathway during injury remains to be dissected. Excessive increases in shear stress (for example, as a result of acute loss of liver mass) could be detrimental, leading to suboptimal liver regeneration. Lorenz *et al.* also did not directly assess whether lack of biomechanical activation of integrin $\beta 1$ and VEGFR3, as might occur in diseases such as diabetes, would lead to decreases in the liver's regenerative potential^{1,2}.

In future, the ideal magnitude of cyclic stretch or shear stress required to initiate the physiological induction of angiocrine factors should be studied. The recruitment of circulating endothelial progenitor cells (EPCs), which are thought to supply the liver with HGF, could also be affected by shear-dependent activation of LSECs, further altering the liver's supply of angiocrine factors¹⁰. Indeed, how increased biomechanical forces alter the delivery of regenerative modulators to the liver, including circulating EPCs, inflammatory cells and platelets, to drive liver growth without encouraging scarring, needs further investigation.

Exactly how do integrin $\beta 1$ and VEGFR3 upregulate angiocrine factors? It is plausible that fluid shear stress induces integrin-mediated nuclear localization of specific transcription factors and so promotes the expression of angiocrine-factor genes^{2–4}. Furthermore, integrin-mediated modulation of the elasticity of the extracellular matrix around hepatocytes in response to shear stress could also modulate hepatocyte proliferation. But what about VEGFR3? Proteins of the VEGFR family are activated by phosphorylation. Biomechanically independent phosphorylation of VEGFR2 on LSECs activates the protein AKT, which recruits the transcription factor Id1 to DNA, inducing the expression of *Wnt2* and *HGF* genes². But the mechanism by which phosphorylation of VEGFR3 turns on angiocrine factors is unknown.

These questions notwithstanding, Lorenz and colleagues' work takes into consideration the complexity of the biophysical environment to which LSECs are exposed *in vivo*, and so solves a mystery that has puzzled liver biologists for decades. The development of strategies that precisely regulate the magnitude of shear stress and cyclic stretch in the liver vasculature might

restore angiocrine-dependent regenerative functions of the liver in pathological conditions, such as in cirrhosis, hepatitis and vascular abnormalities. This could in turn open the door to more-effective therapeutic liver regeneration. ■

Sina Y. Rabbany and Shahin Rafii are in the DeMatteis School of Engineering and Applied Science, Hofstra University, New York, New York 11548, USA, and in the Division of Regenerative Medicine, Ansary Stem Cell Institute, Weill Cornell Medicine, New York. e-mails: sina.y.rabbany@hofstra.edu; s.rafi@med.cornell.edu

MICROBIOLOGY

The electrifying energy of gut microbes

Some bacteria make energy in a process that is accompanied by transfer of electrons to a mineral. A previously unknown electron-transfer pathway now reveals an energy-generation system used by bacteria in the human gut. [SEE LETTER P.140](#)

LATY A. CAHOON & NANCY E. FREITAG

The ability of certain bacteria to transfer electrons has been exploited for a variety of energy-generating applications, such as microbial fuel cells¹, because the flow of charge carried by electrons underlies the process that generates electricity. It was thought that the capacity to achieve substantial levels of electron transfer occurred only in a specialized subset of bacteria. These microbes make energy by a mechanism that requires minerals for the electron-transfer process that accompanies energy generation². On page 140, Light *et al.*³ report the discovery of an electron-transfer pathway in gut bacteria, and reveal that components of this pathway are present in diverse microbial species.

The molecule ATP provides the fundamental energy 'currency' for most cells, and is mainly produced by two mechanisms: fermentation, an anaerobic process in which ATP is generated from a limited repertoire of carbon sources, and respiration, a process that provides a high yield of ATP from a wide array of carbon sources and requires a compound that can accept electrons. In multicellular organisms, respiration involves electron transfer along an electron-transport chain that culminates in electrons being transferred to oxygen⁴.

By contrast, microbes can use a number of alternatives to oxygen as electron acceptors that enable respiration in anaerobic environments lacking fermentable energy sources^{2,5}. For example, the bacteria *Shewanella oneidensis*

1. Rafii, S., Butler, J. M. & Ding, B.-S. *Nature* **529**, 316–325 (2016).
2. Ding, B.-S. *et al.* *Nature* **468**, 310–315 (2010).
3. Hu, J. *et al.* *Science* **343**, 416–419 (2014).
4. Rocha, A. S. *et al.* *Cell Rep.* **13**, 1757–1764 (2015).
5. Lorenz, L. *et al.* *Nature* **562**, 128–132 (2018).
6. Rabbany, S. Y., Ding, B.-S., Larroche, C. & Rafii, S. in *Mechanical and Chemical Signaling in Angiogenesis* (ed. Reinhart-King, C. A.) 19–45 (Springer, 2012).
7. Song, Z. *et al.* *Semin. Cell Dev. Biol.* **71**, 153–167 (2017).
8. Wang, J. H.-C., Goldschmidt-Clermont, P., Wille, J. & Yin, F. C.-P. *J. Biomech.* **34**, 1563–1572 (2001).
9. Baeyens, N., Bandyopadhyay, C., Coon, B. G., Yun, S. & Schwartz, M. A. *J. Clin. Invest.* **126**, 821–828 (2016).
10. DeLeve, L. D. *J. Clin. Invest.* **123**, 1861–1866 (2013).

This article was published online on 26 September 2018.

and *Geobacter metallireducens* reside in mineral-rich environments, and these highly studied microbes have an anaerobic respiration process that uses minerals, such as iron(III) oxide (Fe_2O_3), as respiratory electron acceptors². However, because insoluble mineral deposits cannot be transported into the cell, mineral-respiring bacteria use a mechanism² called extracellular electron transfer (EET), in which electrons are transferred to the exterior of the cell. In the case of these bacteria, this process involves electron transfer from an NADH molecule to components that include a quinone molecule in the lipid membrane and a series of proteins containing haem groups that provide a path for electron transfer. The loss of an electron converts NADH to NAD^+ , which is used in the energy-generation process.

The food-borne bacterial pathogen *Listeria monocytogenes* sometimes has a host-associated part of its life cycle. This bacterium can infect humans, and can proliferate in nutrient-rich environments that enable the use of fermentation as a metabolic strategy⁶. However, although *L. monocytogenes* has a life cycle in which neither minerals nor respiration is crucial for survival, Light *et al.* report that, when *L. monocytogenes* was placed in an electrochemical chamber in which an electrode can trap electrons, an electric current was generated, suggesting that this type of bacterium has the capacity for EET. This report now clarifies evidence presented decades ago⁷, indicating that this bacterium can change extracellular iron in the Fe^{3+} form to the Fe^{2+}

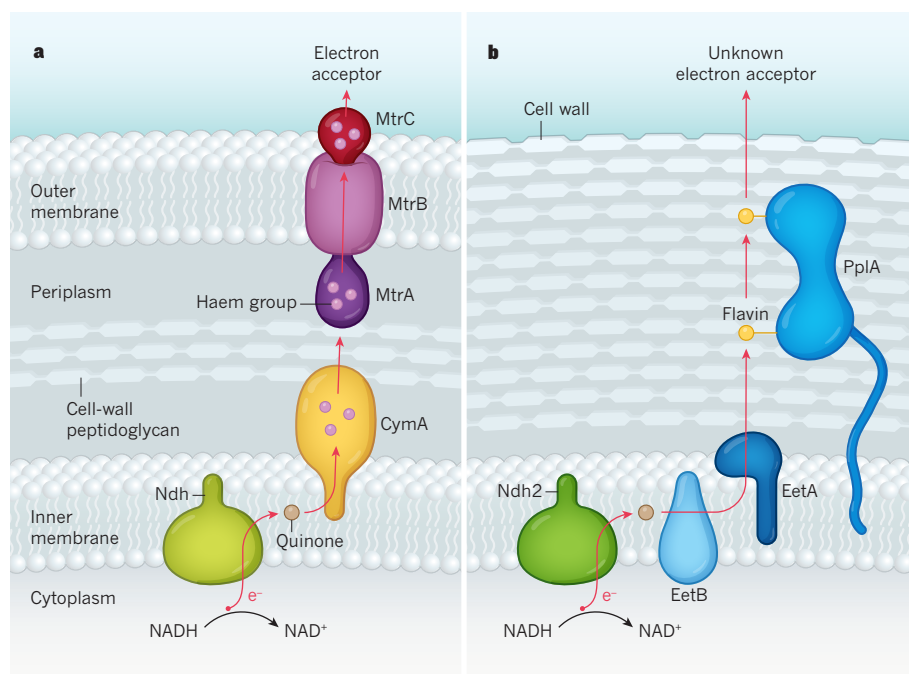


Figure 1 | Bacterial electron-transfer pathways. **a**, In the Gram-negative bacterium *Shewanella oneidensis*, the mechanism that generates energy from the molecule NAD⁺ is accompanied by a process in which an electron (e⁻) from the molecule NADH is transferred outside the cell to a mineral such as iron(III) oxide that acts as an electron acceptor. In this process, known as extracellular electron transfer (EET), the electron-transfer path (red arrow) occurs across two lipid membranes and across the periplasm region, which contains cell-wall material that includes the sugar peptidoglycan. The electron-transfer route towards the cell exterior after crossing the protein Ndh includes a quinone molecule, haem groups associated with the proteins CymA, MtrA and MtrC, and transfer through the protein MtrB. **b**, Light *et al.*³ report a previously unknown EET mechanism in *Listeria monocytogenes*, a Gram-positive bacterium, which has only a single membrane. The authors identified components of this EET system, including the proteins Ndh2, EetB, EetA and PplA (which is associated with two flavin molecules). This newly identified EET process might occur in diverse bacteria, including those in the human gut. The electron acceptor for the pathway is unknown.

form, an alteration that might indicate electron transport out of the cell.

Using a combination of genetic and biochemical approaches, Light *et al.*, true to the name, shed light on the molecular basis of this newly discovered form of EET. They identified the proteins Ndh2, EetB, EetA and PplA as being key components of this process. They show that the initial electron-transfer steps of EET in *L. monocytogenes* resemble those already known in mineral-respiring specialists. For example, electron transfer from the cell cytoplasm to a quinone molecule in the lipid membrane is similar to the steps of a conventional electron-transport chain. However, beyond this point, the mechanisms become more distinct. *L. monocytogenes* is a Gram-positive bacterium, which means that it has a single lipid membrane and a thick cell wall. By contrast, *S. oneidensis* and *G. metallireducens* are Gram-negative bacteria, which have two lipid membranes separated by a region called the periplasm that contains cell-wall material. In these bacteria, tens of haem molecules bound to three types of protein establish a path for electrons to move across the periplasm and the outer lipid membrane⁸. By contrast, in *L. monocytogenes*, a single protein called PplA that contains two flavin molecules suffices to

enable electrons to exit the membrane to reach the cell's exterior (Fig. 1).

Light and colleagues analysed the distribution of the genes for this newly identified EET pathway in the genomes of different bacterial species, and provide evidence of EET activity in species other than *L. monocytogenes* using an electrochemical chamber. They reveal that EET activity occurs in an environmentally and evolutionarily diverse subset of Gram-positive bacteria, most notably in certain bacteria found in the human gut, such as those of the genus *Lactobacillus*.

This observation is intriguing because EET usually provides energy in anaerobic conditions, and growth strategies for such conditions can be important for microbial proliferation in the mammalian gut⁹. Indeed, Light *et al.* found that genes encoding components of the EET system they identified are required for *L. monocytogenes* to grow in anaerobic conditions. Moreover, when the authors monitored the ability of *L. monocytogenes* strains to colonize the mouse gut, the strains deficient in components of this EET system were at a competitive disadvantage, suggesting that EET has a key role in bacterial survival in this context. Investigating the role of EET in host–microbe interactions could

offer an exciting direction for future research.

A central question raised by these findings is why EET might have evolved outside the context of mineral-respiring specialists. The bacterial environment may provide a clue. When microbes such as *L. monocytogenes* live in a host gut, they are immersed in nutrients, including flavin molecules, and Light *et al.* show that the presence of flavins potentially enhances EET activity. The electron-transfer apparatus is simpler in Gram-positive bacteria than in Gram-negative bacteria. It stands to reason that an abundance of environmental flavins might produce a scenario in which evolution favours the minimal investment in protein infrastructure needed to enable EET in certain Gram-positive bacteria. EET might be used by certain mineral-respiring bacteria because it is crucial for their survival, whereas *L. monocytogenes* might use EET because it provides an opportunity to easily generate energy in certain environments.

The electron acceptor used by *L. monocytogenes* for EET is unknown. The bacterium might encounter conditions in which minerals represent an attractive electron acceptor, but it seems more probable that the highly reactive flavins in this pathway aid electron transfer to compounds such as organic soil components, disulfide groups on proteins or even other microbes^{10,11}. If this is the case, in contrast to EET associated with specialized mineral respiration, the EET in *L. monocytogenes* might provide a more flexible mechanism for moving electrons to a variety of environmental acceptors.

It is a shock to the system to consider that microbes might be living highly charged lives in our gut. Light and colleagues' work provides a foundation for future investigation regarding such microbial existence. Furthermore, the characterization of this previously unknown EET mechanism might create opportunities for the design of bacteria-based energy-generating technologies. ■

Laty A. Cahoon and Nancy E. Freitag
are in the Department of Microbiology and Immunology, University of Illinois at Chicago, Chicago, Illinois 60612, USA.
e-mail: nfreitag@uic.edu

1. Lovley, D. R. *Annu. Rev. Microbiol.* **66**, 391–409 (2012).
2. Shi, L. *et al. Nature Rev. Microbiol.* **14**, 651–662 (2016).
3. Light, S. H. *et al. Nature* **562**, 140–144 (2018).
4. Richardson, D. J. *Microbiology* **146**, 551–571 (2000).
5. Glasser, N. R., Saunders, S. H. & Newman, D. K. *Annu. Rev. Microbiol.* **71**, 731–751 (2017).
6. Freitag, N. E., Port, G. C. & Miner, M. D. *Nature Rev. Microbiol.* **7**, 623–628 (2009).
7. Deneer, H. G. & Boychuk, I. *Can. J. Microbiol.* **39**, 480–485 (1993).
8. Nealson, K. H. & Rowe, A. R. *Microb. Biotechnol.* **9**, 595–600 (2016).
9. Winter, S. E. *et al. Science* **339**, 708–711 (2013).
10. Scheller, S., Yu, H., Chadwick, G. L., McGlynn, S. E. & Orphan, V. J. *Science* **351**, 703–707 (2016).
11. Summers, Z. M. *et al. Science* **330**, 1413–1415 (2010).

This article was published online on 12 September 2018.

MEDICAL RESEARCH

Neighbourhood deaths switch cancer subtype

How the same type of cell can form different kinds of tumour isn't always clear. The discovery that cancer subtype in mice is influenced by the type of cell death occurring in the microenvironment provides some insight. SEE ARTICLE P.69

ELI PIKARSKY

The characteristics used to classify tumours, such as the appearance of cancer cells under a microscope, usually reflect the type of cell from which the cancer originated. Yet sometimes the same type of cell can give rise to cancers that are substantially different in appearance and prognosis. The mechanisms that dictate this type of diversity in cancer development are mainly unknown. Seehawer *et al.*¹ reveal on page 69 that the same type of cell can give rise to different types of cancer depending on the sort of cell death that occurs nearby in the tumour microenvironment. This suggests that nearby injury or damage can affect the identity of a cancer.

Human tumour samples are classified in the clinic using a microscope-based technique called histology to assess the shape and form of tumour cells. A prognosis is determined and treatment decisions are made on the basis of this classification. This approach assumes that cancer cells' appearance reflects that of their cell of origin.

A key mechanism that enables tumour cells to retain the characteristics of their founder cell is the formation of heritable types of alteration, known as epigenetic changes. These are chemical modifications, such as the attachment of methyl groups, that are made to DNA and to the histone proteins that associate with it to form chromatin. Epigenetic modifications affect chromatin structure and can have long-term effects on gene expression without changing the genome sequence.

Differences between particular subtypes of tumour can arise if tumours originate from different types of cell residing in normal tissue^{2,3}. But in some cases, the same type of cell can give rise to two different tumour subtypes. One explanation for such a divergence is the presence of mutations that affect cellular appearance⁴. Yet the mechanisms that underlie diversity in cancer subtype are mainly unknown, which is important medically because tumour identity is linked to prognosis and treatment options.

Liver cancer is the second-highest cause of cancer mortality globally⁵, and two common histologically distinguishable subtypes are called hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC).

Originally, it was thought that differentiated hepatocytes, the main type of cell in the liver, give rise to HCC, and that bile-duct cells (also called cholangiocytes) in the liver give rise to ICC. Yet studies in mice indicate that both HCC and ICC can arise from hepatocytes^{6,7}. But how can the same type of cell form two different tumour subtypes that have striking differences in form and progression?

Seehawer and colleagues made a fortuitous discovery when generating liver cancer in mice by the *in vivo* delivery of identical cancer-promoting genes. The animals developed either HCC or ICC, depending on whether the gene-delivery technique was tail-vein injection or electroporation, respectively (Fig. 1). The authors recognized that investigating this observation might shed light on a fundamental aspect of cancer development.

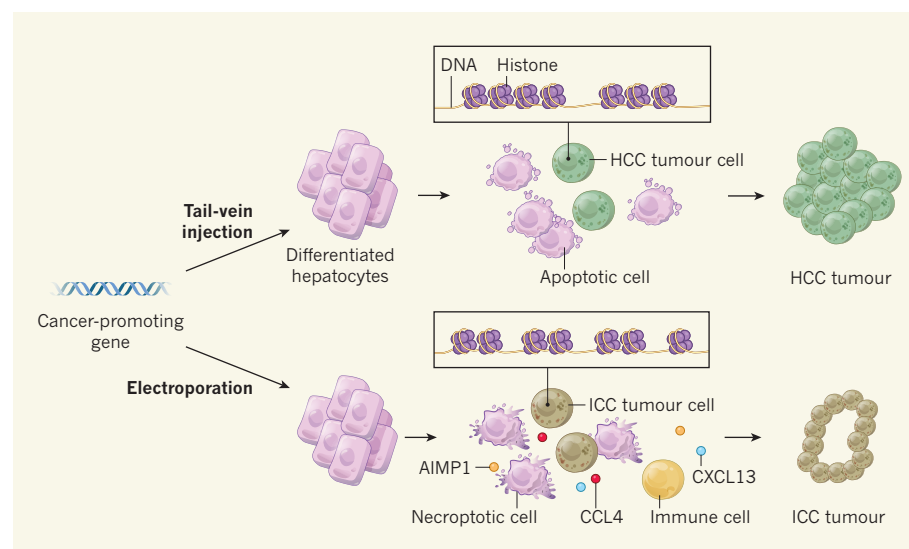


Figure 1 | Nearby cell death influences cancer subtype. Seehawer *et al.*¹ used two techniques (tail-vein injection or electroporation) to transfer the same cancer-promoting genes into differentiated hepatocytes (the main liver cell type) in mice. If tail-vein injection was used, some cells in the microenvironment of the developing tumour underwent a form of cell death called apoptosis, and the tumour subtype that arose was hepatocellular carcinoma (HCC). If electroporation was used, a form of cell death called necroptosis occurred in the microenvironment. Necroptosis was associated with high levels of immune-signalling molecules, including AIMP1, CCL4 and CXCL13, and this form of cell death is generally associated with inflammation and the presence of immune cells. The tumour subtype that arose in this context is called intrahepatic cholangiocarcinoma (ICC). The authors report that there were differences between HCC and ICC cells in the structure of the complex of histone proteins and DNA called chromatin. Such differences can have a long-term effect on gene expression. If cell-death processes affect signalling pathways in nearby tumour-initiating cells in a way that influences chromatin, this might explain the mystery of how the same type of cell can give rise to tumour subtypes that have different appearances and prognoses in humans.

requires a signalling input that is influenced by necroptosis.

Necroptosis is a more inflammatory form of cell death than apoptosis. The authors analysed immune-signalling molecules called cytokines in the livers of mice given the cancer-inducing treatments, and observed differences in the cytokine pattern depending on whether HCC or ICC developed. For example, they noted an increase in the expression of cytokines AIMP1, CCL4 and CXCL13 associated with ICC, compared with their levels in HCC. This raised the question of whether necroptosis-associated inflammation might induce epigenetic changes in hepatocytes that are poised to become cancer cells. The authors found differences in chromatin structure in selected regions of the genome between the two cancer subtypes, although when and how these differences arose is unknown.

Seehawer and colleagues also observed that the transcription-factor protein TBX3 was more highly expressed in HCC than in ICC, whereas the transcription factor PRDM5 had the opposite pattern of higher expression in ICC than in HCC. These expression differences were associated with epigenetic differences in the chromatin structure of the genes encoding these proteins. Remarkably, when the authors analysed samples of human HCC or ICC, they also observed the same pattern of higher TBX3 expression in HCC than in ICC, and higher PRDM5 expression in ICC than in HCC.

Altogether, the evidence suggests that the cell-death conditions prevailing in the liver at the earliest stages of tumour formation might account for the formation of these two different tumour subtypes. Early events in tumour formation are long over by the time a biopsy sample is taken from a human liver; this mouse study could help to illuminate key events that shape the initial steps of tumour formation and result in different cancer identities.

The study provides strong evidence that the tumour microenvironment provides yet-to-be identified signals that can impart long-lasting changes to the fate of cells poised to form cancer cells. Perhaps cytokines might be the drivers of this cancer-subtype switch. If so, the release of such cytokines because of tissue damage or disease might shape the identity of a cancer that is starting to form nearby.

In Seehawer and colleagues' experimental system, the cancer-subtype switch occurred after the acquisition of cancer-promoting genes. However, in human cancers, a different sequence of events is often noted that is thought to occur before the acquisition of cancer-promoting genetic changes. In this scenario, an increased risk of cancer is associated with a process called metaplasia, in which one type of differentiated cell reversibly switches to a different type of differentiated cell. How a predisposition to malignancy arises because of metaplasia is unknown.

It has been reported⁸ that the abnormal accumulation of bile-duct-like structures

in the livers of mice that have chronic liver disease is not due to the proliferation of bile-duct cells, as was previously thought, but that these structures arise from hepatocytes, and that this therefore constitutes a form of metaplasia. Interestingly, a similar pattern of growth of bile-duct-like cells is often observed in human livers^{8,9}. Although this is not usually considered as a form of metaplasia, it is associated with an increase in the risk of liver cancer⁹. Could the switch between HCC and ICC in mice be similar to the process that occurs in liver metaplasia in humans? If it is, Seehawer and colleagues' work might provide insight into how metaplasia increases cancer risk. ■

Eli Pikarsky is at the Lautenberg Center for Immunology and Cancer Research,

IMRIC, and in the Department of Pathology, The Hebrew University of Jerusalem, Jerusalem 9112102, Israel.
e-mail: peh@hadassah.org.il

1. Seehawer, M. *et al.* *Nature* **562**, 69–75 (2018).
2. Van Keymeulen, A. *et al.* *Nature* **525**, 119–123 (2015).
3. Koren, S. *et al.* *Nature* **525**, 114–118 (2015).
4. Yang, J. *et al.* *Cell* **117**, 927–939 (2004).
5. Llovet, J. M. *et al.* *Nature Rev. Dis. Primers* **2**, 16018 (2016).
6. Fan, B. *et al.* *J. Clin. Invest.* **122**, 2911–2915 (2012).
7. Sekiya, S. & Suzuki, A. *J. Clin. Invest.* **122**, 3914–3918 (2012).
8. Tarlow, B. D. *et al.* *Cell Stem Cell* **15**, 605–618 (2014).
9. Ziol, M. *et al.* *Gastroenterology* **139**, 335–343 (2010).

This article was published online on 12 September 2018.

IMMUNOLOGY

Put to sleep by immune cells

The sleep disorder narcolepsy is linked to immune-system genes and is caused by the loss of neurons that express the protein hypocretin. Hypocretin-targeting immune cells have now been found in people with narcolepsy. [SEE ARTICLE P.63](#)

ROLAND S. LIBLAU

The events that lead to the sleep disorder narcolepsy are a long-standing mystery. On page 63, Latorre *et al.*¹ reveal that people with narcolepsy have unusually high levels of a type of immune cell called a T cell, which targets proteins normally present in neurons in the brain. This finding raises the question of whether narcolepsy arises because T cells unleash an autoimmune response against neurons that are important for sleep regulation.

Narcolepsy affects around 1 in 2,000 people². The symptoms usually begin in adolescence or early adulthood, and include daytime sleepiness and, in some cases, cataplexy — sudden muscle weakness during wakefulness that causes falls. A small population of neurons in the brain produces a protein called hypocretin, which controls sleep–wake cycles³, and narcolepsy-like symptoms occur in animals that have defects in genes required for the production of or response to hypocretin⁴. Narcolepsy type 2 is associated with daytime sleepiness, and this can progress to narcolepsy type 1, which is characterized by sleepiness and cataplexy. People with narcolepsy type 1 have abnormally low numbers of hypocretin-producing neurons⁵.

Hypocretin levels in the cerebrospinal fluid that bathes the brain and spinal cord can be measured to help diagnose⁶ narcolepsy type 1,

and such tests provide a way of indirectly monitoring the loss of hypocretin-producing neurons over time. The trajectory of this neuronal loss remains to be fully understood, but can take months or years.

Studies of human genetics have implicated immune-system genes in narcolepsy. Yet whether the immune system contributes to the demise of hypocretin-producing neurons, and if so, how, was unknown. HLA genes encode proteins that can present protein fragments called antigens to T cells, and this interaction can trigger an immune response against cells that contain the specific antigen. Autoimmune diseases are often associated with HLA genes⁷. A version of one such gene, called

“Do T cells that target neuronal proteins other than hypocretin have a role in narcolepsy?”

*HLA-DQB1*06:02*, is present in more than 98% of people with narcolepsy⁸, but it is found in only 15–30% of the general population, depending on ethnicity^{8,9}. Moreover, previous reports^{10,11} suggest that antibodies targeting neuronal proteins are present at a higher than usual frequency in people with narcolepsy.

Latorre and colleagues used various techniques to identify human T cells that recognize specific antigens. The authors tested whether T cells that recognize antigens from hypocretin were present in blood samples

from 19 people with narcolepsy (14 of whom had *HLA-DQB1*06:02*) and from a control group of 13 people who didn't have narcolepsy and carried the *HLA-DQB1*06:02* gene variant. The authors found that all of those with narcolepsy had a type of T cell called a CD4⁺ memory T cell that gave a substantial response to peptide fragments of hypocretin, and this response included the expression of immune-signalling molecules called cytokines (Fig. 1). However, only three members of the control group had T cells that responded to hypocretin, and this response was much weaker than that of the narcolepsy group. When the authors analysed samples of CD4⁺ T cells, they found that the proportion of these cells that recognize hypocretin was more than ten times higher in individuals with narcolepsy than in control individuals. This increased T-cell reactivity to hypocretin has also been reported independently¹², and strongly suggests that autoimmunity has a role in narcolepsy.

The authors' in-depth *in vitro* analysis of the samples from people with narcolepsy enabled the specific hypocretin peptides recognized by the T cells to be characterized, together with the versions of the T-cell receptors that were involved in antigen recognition. Unexpectedly, most of these T cells did not recognize hypocretin peptides bound to the HLA-DQ6 proteins that are encoded by *HLA-DQB1*06:02*. Instead, HLA proteins called HLA-DR were involved.

Several explanations for this are possible. Maybe an initial T-cell response to hypocretin-producing neurons is dominated by T cells that recognize antigens bound to HLA-DQ6, but over time the response shifts to T cells that recognize antigens presented on HLA-DR. An observation consistent with this model is the recent report¹² that the proportion of T cells that respond to hypocretin presented by HLA-DQ6 is higher than normal in people with recent-onset narcolepsy. Another possibility is that T cells in the bloodstream (tested by the authors) have different HLA-binding preferences from those in the brain. Or perhaps *HLA-DQB1*06:02* might help to promote the development of a repertoire of T cells that recognize antigens in association with HLA-DR proteins¹³.

The authors made the puzzling finding that immune cells called antigen-presenting cells, which, as their name indicates, can present antigens to T cells, did not cleave hypocretin into the specific peptides to which the T cells responded. This suggests that hypocretin needs to be cleaved, either extracellularly or in neurons that express hypocretin, to enable presentation of these peptides by antigen-presenting cells.

Previous epidemiological studies^{14,15} have revealed an increased risk of narcolepsy in people who were vaccinated against the 2009 H1N1 strain of influenza virus. Crossreactivity can occur if a T cell recognizes antigens from

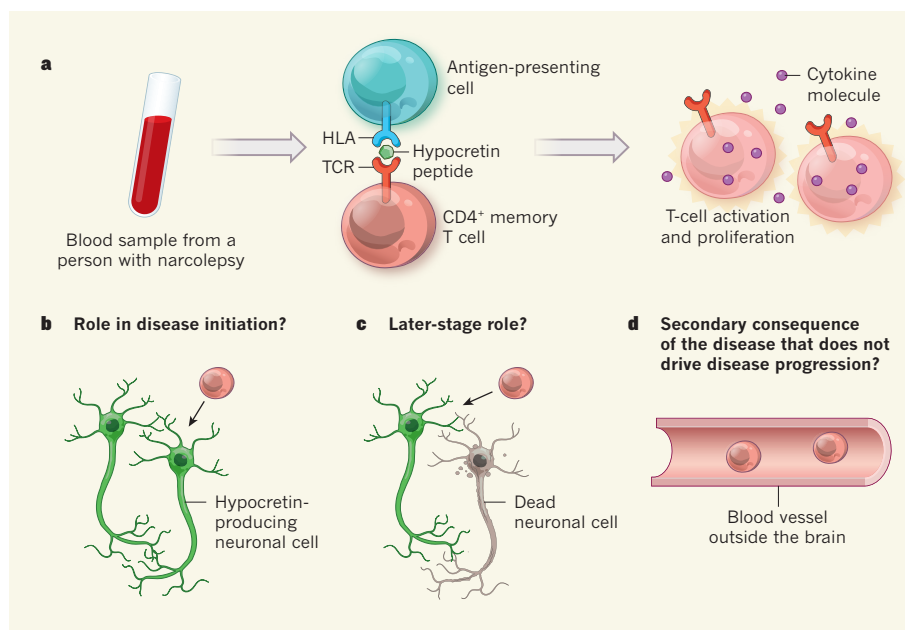


Figure 1 | Immune cells and narcolepsy. The sleep disorder narcolepsy is caused by loss of brain cells that produce the protein hypocretin. **a**, Latorre *et al.*¹ analysed blood samples from 19 people with narcolepsy. They stimulated a type of immune cell in the samples, called a CD4⁺ memory T cell, with antigen-presenting cells, which had an HLA protein displaying a peptide fragment of hypocretin protein on their surfaces. In all the samples, this type of T cell recognized the antigen when the antigen was bound to the T-cell receptor (TCR). This recognition led to T-cell proliferation and activation, and the production of immune-stimulating molecules called cytokines. Whether and, if so, how such hypocretin-targeting T cells contribute to narcolepsy is unknown. **b**, One possibility is that they have a role in the processes that cause disease onset and aid the initial killing of hypocretin-producing neurons. **c**, Or perhaps these T cells contribute to the processes that kill neurons as the disease progresses. **d**, Another possibility is that the presence of these T cells is a secondary consequence of the disease (if, for instance, they are generated in an immune response induced by neuronal death), but these T cells do not contribute to disease progression because, for example, they do not have access to hypocretin-producing neurons.

different types of protein that have some structural similarity¹⁶. Latorre *et al.* report that the T cells they tested that can recognize hypocretin do not recognize antigens from the H1N1 flu virus, and thus do not show crossreactivity. The nature of the link between vaccination against the 2009 H1N1 flu strain and the narcolepsy cases that arose post-vaccination remains unknown.

There is a dearth of information about the immunological processes that occur in the brain during narcolepsy. Latorre and colleagues analysed samples of cerebrospinal fluid from seven people with narcolepsy. One of them had hypocretin-recognizing T cells of a type termed CD8⁺. Further studies are therefore needed to determine whether T cells that recognize hypocretin are enriched in the cerebrospinal fluid of people who have narcolepsy.

Latorre and colleagues noted the presence of T cells that recognize hypocretin in blood samples of two people with narcolepsy type 2 who lacked *HLA-DQB1*06:02*, and in whom hypocretin levels were not abnormally low. This challenges the model that targeting of hypocretin by such T cells is involved in neuronal destruction. However, the authors did detect a high frequency of hypocretin-specific CD8⁺ T cells in blood samples from a person with

narcolepsy type 2 who subsequently developed cataplexy. Therefore, it is possible that this person's hypocretin-producing neurons had been undergoing a process of destruction. This observation is consistent with results from a mouse study showing that CD8⁺ T cells that recognize an antigen expressed by hypocretin-producing neurons can directly kill these neurons¹⁷.

Several questions remain to be answered. For example, do T cells that target neuronal proteins other than hypocretin have a role in narcolepsy? What are the contributions of CD4⁺ and CD8⁺ T cells, and are other immune cells or antibodies involved? CD8⁺ T cells recognize antigens bound to HLA proteins that belong to a group called HLA class I, which can be present on neurons. However, if CD4⁺ T cells target neurons, it is unclear how this occurs, because neurons do not express HLA class II proteins (a group that includes HLA-DQ6 and HLA-DR) that CD4⁺ T cells bind to¹⁸.

Firmly establishing whether there is a causal relationship between the presence of hypocretin-targeting T cells and narcolepsy is a key issue, because if this is confirmed it would provide a therapeutic target for narcolepsy. Some insights might come from studying immune responses over time in



50 Years Ago

A chain of ecological events extraordinary for the British Isles is reported in this week's issue of *Nature* ... They concern the unusual bloom of a planktonic alga, the dinoflagellate *Gonyaulax tamarensis*, off the Northumbrian coast of Britain in May this year. This eventually led to the illness of more than eighty people through mussel poisoning and the deaths of about 80 per cent of the breeding population of shags on the Farne Islands ... Fortunately for the mussel eaters of Northumberland, most of the mussels eaten had been well cooked, the juices had been drained away, and in this way some of the poison was eliminated. Otherwise there would probably have been much more serious effects and even some deaths.
From *Nature* 5 October 1968

100 Years Ago

In North Wales, on August 20 ... I saw a rainbow-effect which was quite new to me. The summit of Tryfaen ... has three sharp, rocky peaks ... We had climbed up the eastern cliff in a south-westerly gale, which brought up much cloud with some light showers, and were sitting just below the top of the southern peak. The Holyhead road lay north-east and 2000 ft. below us. From it rose the upright portion of a brilliant rainbow. At the centre of its circle was the shadow of our peak with those of the other two peaks to the left of it, all sharply defined. Around the shadow of our peak was a most vivid and persistent bow, the smallest I have ever seen, the radius of the inner edge being about half that of the outer ... Outside this bow (which had the colours in regular rainbow order, red outside) was part of a third bow of perhaps double the diameter, but dim and intermittent.

From *Nature* 3 October 1918

patients, ideally starting at disease onset.

It would indeed be valuable to obtain a precise picture of the contribution of hypocretin-specific CD4⁺ T cells to narcolepsy. Do these cells initiate the disease-causing process? Do they contribute to the later-stage progression? Are they present as a secondary consequence of the disease but do not contribute to disease progression? There are many possibilities to consider. For example, do unidentified immune cells cause the demise of hypocretin-producing neurons and lead to a second wave of immune cells, such as hypocretin-specific T cells, that target the neurons? Or perhaps such second-wave T cells might not contribute to the progression of narcolepsy at all, because they might not interact with hypocretin-producing neurons.

If further experiments strengthen the proposed link between increased T-cell reactivity to hypocretin and neuronal damage, Latorre and colleagues' study might lead to targeted immune therapies. If this is the case, such treatments would probably be developed to target the immune system at the time of onset of narcolepsy. ■

Roland S. Liblau is at the Center for Physiopathology of Toulouse Purpan, University of Toulouse, CNRS, INSERM, Toulouse III University, Toulouse 31024, France. e-mail: roland.liblau@inserm.fr

1. Latorre, D. *et al.* *Nature* **562**, 63–68 (2018).
2. Scammell, T. E. N. *Engl. J. Med.* **373**, 2654–2662 (2015).
3. de Lecea, L. *et al.* *Prog. Brain Res.* **198**, 15–24 (2012).
4. Sakurai, T. *Curr. Opin. Neurobiol.* **23**, 760–766 (2013).
5. Liblau, R. S., Vassalli, A., Seify, A. & Tafti, M. *Lancet Neurol.* **14**, 318–328 (2015).
6. Dauvilliers, Y., Arnulf, I. & Mignot, E. *Lancet* **369**, 499–511 (2007).
7. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. *Nature Rev. Immunol.* **18**, 325–339 (2018).
8. Behalf of the European Narcolepsy Network (EU-NN). *Sleep* **37**, 19–25 (2014).
9. Mignot, E. *et al.* *Am. J. Hum. Genet.* **68**, 686–699 (2001).
10. Cvetkovic-Lopes, V. *et al.* *J. Clin. Invest.* **120**, 713–719 (2010).
11. Ahmed, S. S. *et al.* *Sci. Transl. Med.* **7**, 294ra105 (2015).
12. Luo, G. *et al.* Preprint at bioRxiv <https://doi.org/10.1101/378109> (2018).
13. Sharon, E. *et al.* *Nature Genet.* **48**, 995–1002 (2016).
14. Nguyen, X. H., Saoudi, A. & Liblau, R. S. *Curr. Opin. Neurol.* **29**, 362–371 (2016).
15. Sarkanen, T. O., Alakuijala, A. P. E., Dauvilliers, Y. A. & Partinen, M. M. *Sleep Med. Rev.* **38**, 177–186 (2018).
16. Birnbaum, M. E. *et al.* *Cell* **157**, 1073–1087 (2014).
17. Bernard-Valnet, R. *et al.* *Proc. Natl Acad. Sci. USA* **113**, 10956–10961 (2016).
18. Liblau, R. S., Gonzalez-Dunia, D., Wiendl, H. & Zipp, F. *Trends Neurosci.* **36**, 315–324 (2013).

The author declares competing financial interests. See go.nature.com/2nsa0ol for details.

This article was published online on 12 September 2018.

MATERIALS SCIENCE

Morphing into action

An organic polymer exhibits a phase transition that is associated with improved electromechanical properties. This feature links organic polymers with widely used perovskite materials, and could have many applications. SEE LETTER P.96

RONALD E. COHEN

The properties of a material can change dramatically in the vicinity of a phase transition. For example, the electromechanical properties of ferroelectric materials are greatly enhanced in a transition region known as a morphotropic phase boundary. Here, the material's electric polarization (dipole moment per unit volume) rotates from one direction to another. This phenomenon is well studied in perovskite materials, but on page 96, Liu *et al.*¹ report similar behaviour in an organic polymer. The discovery could open up various applications, such as in medical instruments, power-generating clothes and improved safety devices for structures and vehicles.

Active materials transform energy from one type into another. For example, ferroelectrics convert electrical energy into mechanical energy, and vice versa. Ferroelectrics are used

in technologies ranging from medical ultrasound systems to fuel-injection and crash-sensing equipment in cars. They even have applications in refrigeration and the generation of power from excess heat.

The most commonly used ferroelectric is the perovskite lead zirconate titanate (PZT). Although PZT is cheap, it contains toxic lead and is a hard ceramic that can be produced only at high temperatures. The lead content limits the material's use in applications such as tiny pumps, and motors that are permanently installed in the human body. Single-crystal 'relaxor' ferroelectrics have also been developed. These materials have much higher energy-conversion efficiencies than PZT and are advancing the resolution of medical ultrasound systems, for example. But they are much more expensive than PZT, and also contain lead.

In addition, there are active materials based on lead-free polymers. These materials can

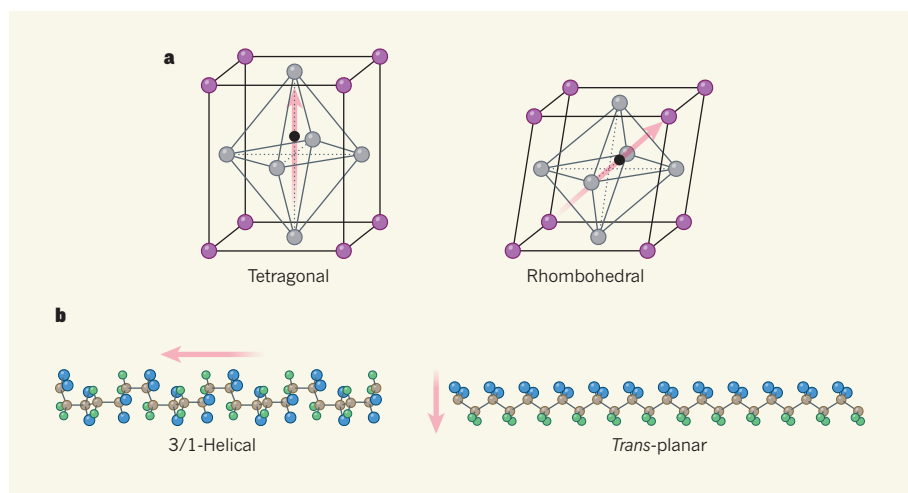


Figure 1 | Two types of phase transition. **a**, The inorganic material lead zirconate titanate (PZT) has different electric polarizations (arrows) on the two sides of a phase transition. Away from the transition region, PZT is tetragonal on the titanium-rich side and rhombohedral on the zirconium-rich side. In the transition region, the polarization rotates between the two different directions, and the material exists in intermediate (monoclinic) phases (not shown). The transition is termed morphotropic because of these shape changes. Purple, lead; grey, oxygen; black, titanium or zirconium. **b**, Liu *et al.*¹ report a similar type of phase transition for the organic polymer system poly(vinylidene fluoride-co-trifluoroethylene) (P(VDF-TrFE)). VDF-poor compositions of this material exist in what is known as a 3/1-helical phase, with a polarization that points along the axis of the polymer chains. VDF-rich compositions are in a *trans*-planar phase, with a polarization that is perpendicular to the chains. For simplicity, structures are shown for chains of VDF, rather than for P(VDF-TrFE). Brown, carbon; blue, fluorine; green, hydrogen.

be produced inexpensively and moulded in a similar way to plastics, but have much lower energy-conversion efficiencies than PZT. Now, Liu and colleagues have demonstrated a way to boost the efficiencies of polymer ferroelectrics, making these materials more competitive with conventional, inorganic ferroelectrics.

The temperature–composition phase diagram of PZT was established in the 1950s². It has an almost-vertical morphotropic phase boundary (MPB), positioned where the ratio of titanium to zirconium is about 50:50. On the titanium-rich side of the boundary, PZT has a tetragonal perovskite structure, whereas on the zirconium-rich side, it has a rhombohedral perovskite structure (Fig. 1a).

The idea of an MPB goes back to the nineteenth century, and is not restricted to perovskites. A paper from 1890 discusses the origin of the term morphotropic³, and suggests that it was coined in 1870 by the mineralogist Paul Heinrich von Groth⁴ — who, incidentally, founded the scientific journal *Zeitschrift für Kristallographie und Mineralogie* and was a professor of mineralogy and curator of minerals in Munich.

The word morphotropic is related to polymorphic. Polymorphs are different crystal structures of compounds that have the same chemical composition — such as graphite and diamond, which are both forms of carbon. A phase boundary between such different polymorphs is not useful for applications of active materials. This is because the phases can exist in a metastable form far from the region of the phase diagram in which they are stable.

Consequently, for example, diamonds can exist on Earth's surface and can even be grown in a metastable form at low pressures.

By contrast, an MPB describes a transition that occurs smoothly, with the crystals changing shape through the transition⁵. Today, MPBs are most commonly thought of in the context of inorganic perovskites. But Groth originally applied the term to organic crystals, such as benzene, as different chemical replacements are being made — for example, switching chlorine atoms with hydrogen atoms⁴. In the past few decades, the importance of morphotropism to organic crystals has been rediscovered and re-emphasized⁶.

Liu and colleagues considered the ferroelectric polymer system poly(vinylidene fluoride-co-trifluoroethylene) (P(VDF-TrFE)), and synthesized systems that had a range of vinylidene fluoride (VDF) contents. They found that VDF-poor compositions existed in what is known as a 3/1-helical phase, whereas VDF-rich compositions were in a *trans*-planar phase (Fig. 1b). The electromechanical properties of the material were maximal in the transition region between these two phases, similar to what happens at the MPB in inorganic perovskites.

Close analogies for P(VDF-TrFE) among perovskites are relaxor ferroelectrics, such as lead magnesium niobate–lead titanate (PMN-PT). In PMN-rich compositions of PMN-PT, and in pure PMN, a ferroelectric phase is not observed because the atoms are displaced from lattice sites in a disordered manner. Similarly, Liu *et al.* saw no ferroelectric phase in VDF-poor compositions of

P(VDF-TrFE) as a result of disorder in the polymer chains.

In the case of perovskites, it was discovered in the past few decades that the simple picture of a boundary between different structural phases needs to be replaced with a transition region in which the materials exist in intermediate (monoclinic) phases^{7,8}. In this region of the phase diagram, the direction of the material's polarization rotates with varying composition, applied electric fields or stresses.

It was also found that varying composition is not required to generate behaviour akin to an MPB, and that even pure lead titanate exhibits such behaviour⁹. More specifically, lead titanate changes shape from tetragonal to rhombohedral through monoclinic intermediates under pressure. In the transition region, theory predicts that the coupling between electrical and mechanical energy will be extremely high — larger than that of known materials at ambient pressure¹⁰.

At first glance, P(VDF-TrFE) might seem about as different from perovskites as possible, but the concept of polarization rotation also holds in this polymer system. In one phase, the polarization points along the axis of the polymer chains, whereas in the other phase, it is perpendicular to the chains. The polarization direction rotates in the MPB region^{11–14}.

Now that behaviour reminiscent of an MPB has been observed in an organic molecular system, the question arises of whether similar behaviour might be induced by applying stress or electric fields, rather than by changing the chemical composition. Moreover, realizing such an effect using strain could enable materials to be easily tuned, giving rise to even more potential applications. ■

Ronald E. Cohen is at the Extreme Materials Initiative, Geophysical Laboratory, Carnegie Institution for Science, Washington DC 20015, USA, the Department of Earth and Environmental Sciences, Ludwig Maximilian University of Munich, Munich, Germany, and the Department of Physics and Astronomy and London Centre for Nanotechnology, University College London, London, UK. e-mail: rcohen@carnegiescience.edu

1. Liu, Y. *et al.* *Nature* **562**, 96–100 (2018).
2. Haertling, G. H. *J. Am. Ceram. Soc.* **82**, 797–818 (1999).
3. Retgers, J. W. Z. *Phys. Chem.* **6U**, 193–236 (1890).
4. Groth, P. *Ann. Phys.* **217**, 31–43 (1870).
5. Everitt, C. *Chemistry in Encyclopaedia Britannica* Vol. 6, 11th edn 74 (Cambridge Univ. Press, 1910).
6. Kálmán, A. *Acta Cryst. B* **61**, 536–547 (2005).
7. Fu, H. & Cohen, R. E. *Nature* **403**, 281–283 (2000).
8. Noheda, B. *et al.* *Phys. Rev. Lett.* **86**, 3891 (2001).
9. Ahart, M. *et al.* *Nature* **451**, 545–548 (2008).
10. Wu, Z. & Cohen, R. E. *Phys. Rev. Lett.* **95**, 037601 (2005).
11. Lovinger, A. J. *Science* **220**, 1115–1121 (1983).
12. Nakhmanson, S. M., Nardelli, M. B. & Bernholc, J. *Phys. Rev. Lett.* **92**, 115504 (2004).
13. Nakhmanson, S. M., Nardelli, M. B. & Bernholc, J. *Phys. Rev. B* **72**, 115210 (2005).
14. Tsutsumi, N., Okumachi, K., Kinashi, K. & Sakai, W. *Sci. Rep.* **7**, 15871 (2017).

A new era in the search for dark matter

Gianfranco Bertone^{1*} & Tim M. P. Tait^{1,2*}

There is a growing sense of ‘crisis’ in the dark-matter particle community, which arises from the absence of evidence for the most popular candidates for dark-matter particles—such as weakly interacting massive particles, axions and sterile neutrinos—despite the enormous effort that has gone into searching for these particles. Here we discuss what we have learned about the nature of dark matter from past experiments and the implications for planned dark-matter searches in the next decade. We argue that diversifying the experimental effort and incorporating astronomical surveys and gravitational-wave observations is our best hope of making progress on the dark-matter problem.

The fall of natural weakly interacting massive particles

The existence of dark matter has been discussed for more than a century^{1,2}. In the 1970s, astronomers and cosmologists began to build what is today a compelling body of evidence for this elusive component of the Universe, based on a variety of observations, including temperature anisotropies of the cosmic microwave background, baryonic acoustic oscillations, type Ia supernovae, gravitational lensing of galaxy clusters and rotation curves of galaxies^{3,4}. The standard model of particle physics contains no suitable particle to explain these observations, and thus dark matter arguably represents a glimpse of physics beyond the standard model. Proposed candidates for dark matter span 90 orders of magnitude in mass, ranging from ultralight bosons (often referred to as ‘fuzzy dark matter’⁵) to massive primordial black holes—a possibility that has received renewed interest after the detection of gravitational waves from the merger of black holes several tens of times more massive than the Sun by the Laser Interferometer Gravitational-wave Observatory (LIGO) and Virgo^{6,7}.

The class of dark-matter candidates that has attracted the most attention over the past four decades is weakly interacting massive particles (WIMPs). WIMPs appeared for a long time as a perfect dark-matter candidate, as new particles at the weak-interaction mass scale (or weak scale; approximately between 10 GeV and 1 TeV) would be produced naturally with the right relic abundance in the early Universe⁸ while possibly alleviating the infamous hierarchy problem⁹, which has been a main driver of particle physics for roughly four decades¹⁰. Despite much effort, no particle other than a standard-model-compatible Higgs boson has been convincingly detected at the weak scale so far—a circumstance that, as long anticipated¹¹, raises the possibility that natural WIMPs may have been nothing more than an attractive red herring¹².

The hierarchy problem is a consequence of the fact that quantum mechanics inevitably mixes up phenomena from all energy scales by allowing virtual particles to participate even in reactions whose energies are far too small to actually produce them. As a result, low energy quantities, such as the Higgs mass, can potentially receive very large corrections from the virtual influence of much heavier particles. The influence of heavy particles is particularly pronounced for scalar bosons such as the Higgs boson and introduces corrections to the effective Higgs mass that are proportional to the masses of the virtual heavy states, so that the effective Higgs mass is the sum of a fundamental intrinsic value plus the correction terms.

Because it is generally expected that new particles will appear at the Planck energy scale, which is associated with quantum gravity,

the observed Higgs mass at the weak scale appears highly unnatural, requiring an incredibly fine-tuned cancellation between the individually much larger intrinsic contribution and the correction terms, such that their sum is the value observed at the Large Hadron Collider (LHC). Natural theories introduce additional particles and symmetries, which are arranged so that these large corrections cancel each other out, protecting the Higgs mass from the influence of heavy mass scales.

The prototypical natural theory is the minimal supersymmetric (SUSY) standard model, which introduces an additional partner for each standard-model particle. In addition, the partners of electroweak bosons are predicted to be WIMPs and thus are natural dark-matter candidates. However, most of the parameter space of natural simple SUSY models is essentially ruled out¹³. Although it is still possible to identify ‘natural’ realizations of SUSY—for example, in regions of the parameter space of the phenomenological minimal SUSY model¹⁴—it is undeniable that null searches are constraining larger and larger portions of the parameter space of SUSY theories, which begs the question of how much fine-tuning one is willing to accept before giving up the hope of discovering SUSY¹⁵.

Alternatives to natural WIMPs

Non-natural WIMPs

As a result of the lack of evidence for supersymmetry, naturalness is beginning to lose its lustre as the guiding principle for constructing theories of physics beyond the standard model. Although the shift away from WIMPs, which arises from extensions of the standard model that address naturalness, is inevitable, WIMPs themselves remain viable dark-matter candidates in an appropriate context. For example, there are types of interaction that lead to highly suppressed indirect and direct signals, although such particles remain accessible to the LHC, provided that their masses are sufficiently small¹⁶. With naturalness removed as the primary guide to theories of WIMPs, such particles evolve into a more general class of particles that achieve the appropriate relic density through self-annihilation.

This wider definition of WIMPs—which is already reflected in the adoption of simplified models¹⁷ and effective field theories¹⁸ in the presentation of collider results—leads to a richer landscape of phenomenology. For example, the range of WIMP masses expands to encompass masses as low as around 1 MeV or as high as around 100 TeV. This wider parameter space demands new kinds of WIMP searches, such as scattering of WIMP-like particles with masses below 1 GeV from electrons¹⁹ or the use of superconductors²⁰, superfluids²¹ or Dirac

¹GRAPPA Institute and Institute of Physics, University of Amsterdam, Amsterdam, The Netherlands. ²Department of Physics and Astronomy, University of California, Irvine, CA, USA.

*e-mail: g.bertone@uva.nl; ttait@uci.edu

materials²². Such light dark-matter particles would have already been observed if their annihilation cross-sections into standard-model particles were large enough to explain their abundance in the Universe. As a result, viable models typically invoke similarly light ‘dark’ force carriers into which the dark matter can annihilate, and which subsequently decay into standard-model states. Because they have small masses and must interact at some level with the standard-model particles, these dark force carriers can be probed using high-intensity, low-energy accelerators²³. Another complementary avenue is the search for teraelectronvolt-energy γ -rays produced in the annihilation of ultraheavy dark-matter particles with the upcoming γ -ray Cherenkov Telescope Array (CTA)^{24,25}.

Axions

Another very popular class of dark-matter candidate is that of axions and axion-like candidates. Quantum chromodynamics (QCD) axions are light, very weakly coupled particles that arise as a byproduct in theories that solve the ‘strong-CP problem’ in QCD. The symmetries of the standard model of particle physics allow the strong nuclear force to include an electric dipole moment for the neutron, which represents an asymmetry in the charge distributions of its constituent quarks. However, measurements indicate that the neutron electric dipole moment is about 10^{-10} times smaller than expected, which necessitates a dynamical explanation. The dynamics that would cancel the neutron electric dipole moment also produces a new particle: the axion²⁶.

Many constraints exist on axions and axion-like models. A class of searches typified by the Axion Dark Matter Experiment (ADMX)²⁷ uses a magnetic field to convert the background of axions on Earth into an electromagnetic signal. Such searches have successfully excluded a window of axion parameter space with masses around 2 meV, and future measurements are expected to probe masses up to about 40 meV. In addition, there is vigorous theoretical activity exploring new ways to probe a wider range of axion masses^{28–30}.

Sterile neutrinos

Another well motivated candidate is the sterile neutrino, which experiences a diluted form of the weak nuclear force through mixing with ‘ordinary’ active neutrinos. Such particles are typically included in theories that explain experiments that have found neutrinos to be massive, in contrast to the predictions of the standard model. Although their residual weak interactions indicate that sterile neutrinos will ultimately decay if both their mass and mixing are small enough, this decay may occur slowly enough so that they remain in the Universe today as a form of dark matter. Such neutrinos can be produced in the early Universe through a variety of different physical mechanisms^{31–34} with an appropriate abundance.

Although the lifetime of a sterile neutrino playing the role of dark matter must be long enough so that the vast majority of such particles have not yet decayed, quantum mechanics dictates that some will decay more rapidly, leading to a source of mono-energetic photons with energy close to half of the neutrino mass. In fact, an unidentified emission line at 3.5 keV in the stacked X-ray spectrum of 73 galaxy clusters might be a hint of the decay of sterile neutrinos³⁵, although debate about the origin of this line is still ongoing³⁶. Future X-ray telescopes, such as eRosita, X-ray Astronomy Recovery Mission (XARM), Athena and Lynx, should help to clarify the origin of this emission³⁷, and future accelerator searches, such as with the Separator for Heavy Ion reaction Products (SHIP), will provide a complementary probe of the relevant parameter space.

No stone left unturned

There is a plethora of other possible explanations for the nature of dark matter (see Fig. 1 for a diagrammatic representation), including fuzzy dark matter (10^{22} eV), gravitationally produced WIMPzillas³⁸, superfluid dark matter³⁹, macroscopic objects such as macros (10^{22} – 10^{24} g)⁴⁰ and primordial black holes ($10M_{\odot}$, where M_{\odot} is the mass of the Sun). Therefore, the new guiding principle should be ‘no stone left unturned’:

we should look for dark matter not only where theoretical predictions dictate that we ‘must’, but wherever we can. Casting a wider theoretical net offers the possibility of discovering new classes of dark-matter candidates and new experimental opportunities to search for them, and also helps assemble a ‘composite image’ of everything that we currently know about the space of possibilities that are consistent with existing measurements.

Probing dark matter with astronomical observations

Departures from the lambda cold dark matter model

Given the current absence of evidence for dark-matter particles from laboratory experiments, it is of utmost importance to extract as much information as possible from astronomical observations. Dark-matter couplings other than that of gravity with itself or with standard-model particles, or a non-negligible velocity dispersion, could lead in principle to measurable differences between observations and lambda cold-dark matter (LCDM) model predictions⁴¹. It is generally important to search for ‘cracks’ in the LCDM model by carefully testing its underlying assumptions and observational predictions. An intriguing example is the discrepancy at the 3.7σ level between cosmological³ and local measurements of the Hubble constant⁴². We stress that systematic errors in observations, or mismodelling of specific physical processes, should not be mistaken for failures of the underlying LCDM model. It is perhaps not a surprise in this sense that most of the claimed problems of standard cosmology, such as the cusp–core, too-big-too-fail and missing-satellites problems⁴¹, arise in the deeply nonlinear regime. Model predictions are in this case based on numerical simulations that encode complex processes, such as stellar formation and supernova and black-hole feedback, by means of an effective ‘sub-grid’ description⁴³, which is by construction a potential source of systematic errors. This should not of course deter us from extensively testing the predictions of standard cosmology by exploiting the wealth of information that will arise from upcoming astronomical surveys—such as those using the Large Synoptic Survey Telescope (LSST), Dark Energy Spectroscopic Instrument (DESI), Euclid and the Wide-Field Infrared Survey Telescope (WFIRST)—while improving the quality and predictive power of numerical simulations.

Self-interactions

A key property of dark matter that astronomical observations might help disproving is its collisionless nature. Dark matter self-interactions might actually help alleviate claimed tensions between numerical simulations and observations at small cosmological scales^{44,45}. We can search for the imprint of dark-matter self-interactions in a number of ways. First, self-interactions can modify the shapes of dark-matter haloes⁴⁴; in fact, they tend to make the central parts of dark-matter haloes more spherically symmetric than expected in collisionless scenarios. By comparing the shape of galaxy clusters in numerical simulations with that inferred from lensing and X-ray observations, it is possible to set an upper limit⁴⁶ on the velocity-independent, elastic cross-sections σ of self-interacting dark matter of mass m : $\sigma/m \approx 1 \text{ cm}^2 \text{ g}^{-1}$. Only very recently the first full simulations of galaxy clusters that incorporate both baryonic processes and dark-matter self-interactions have been obtained⁴⁷. Although much remains to be understood, it is encouraging that these simulations appear to support the analytical models tying the properties of self-interacting dark matter to the observed distribution of baryons⁴⁸.

Second, the trace of dark-matter self-interactions could be found in merging systems such as cluster mergers and minor infalls^{49,50}. The observables in this case would be the offset between the galaxies and the dark matter (in addition to the offset between dark matter and gas) due to the possible non-collisional nature of dark matter⁵¹, and the amount of ‘sloshing’ and ‘wobbling’ of galaxies around the centre of the dark-matter halo^{41,52}. As in the case of halo shapes, it is urgent to further investigate the complex interplay between gas cooling, active-galactic-nuclei feedback and dark-matter physics using full hydrodynamical simulations, and understand the mapping between the



Fig. 1 | Possible solutions to the dark-matter problem. Visualization of the possible solutions to the dark-matter problem in the form of a mind-map diagram. The label ‘little Higgs’ refers to dark-matter candidates that arise in the framework of little Higgs models¹ and ‘extra dimensions’

indicates candidates related to theories with extra space dimensions¹. TeVeS, tensor–vector–scalar theory; MOND, modified Newtonian dynamics; MaCHOs, massive compact halo objects¹.

properties of self-interacting dark matter and observables, in preparation for the wealth of observational data that will arise from upcoming astronomical surveys.

Substructures

A generic key property of dark matter in the standard cosmological model is that it is cold—that is, non-relativistic—at the epoch of structure formation and has a free-streaming length much smaller than the size of galaxies. This implies the existence of a large number of sub-dwarf galaxy dark structures in galactic haloes. If dark matter is warm or, more generally, if its power spectrum is suppressed at small astrophysical scales, then we might identify it by probing the actual number of substructures in the Universe. A powerful probe of the power spectrum at small scales is the Lyman- α forest in the spectra of high-redshift quasars⁵³. This technique allows us to set a 2σ lower limit of 5.3 keV on the warm-dark-matter particle mass⁵⁴ and a 2σ lower limit of 37.5×10^{-22} eV on the mass of fuzzy-dark-matter particles⁵⁵. Observations with the future high-resolution spectrograph of the European Extremely Large Telescope (E-ELT) and with low-resolution, low-signal-to-noise-ratio quasar spectra measured by DESI should allow to substantially improve the current bounds thanks to a larger statistical sample and a better determination of the thermal state of the intergalactic medium.

Another interesting strategy to detect these dark substructures is the search for perturbations induced by sub-dwarf galaxy clumps on cold stellar streams^{56–58}. Thanks to surveys such as Gaia, which is currently taking data, and LSST, it should be in principle possible to detect impacts induced on stellar streams by subhaloes with masses⁵⁹ as low as $10^7 M_\odot$. By analysing the power spectrum of the fluctuations of the stellar density, stream observations might even enable us to probe subhaloes with masses⁵⁸ down to $10^5 M_\odot$. This method should allow us to set stringent constraints on the mass of thermal dark-matter

relics using LSST data, and possibly yield an actual measurement of the dark-matter particle mass if this mass⁶⁰ is of the order of 1 keV. A more direct way of detecting dark-matter substructures is via gravitational lensing. Although dark-matter subhaloes are not compact enough to be detectable, for example, with microlensing searches, they can modify the flux ratio of multiply lensed quasars^{61–64} and are potentially detectable via gravitational imaging, as a perturbation of magnified arcs and Einstein rings⁶⁵. In addition to lens substructures, low-mass dark-matter haloes along the line of sight of the lens can act as perturbers and dominate the signal by an amount that depends on the lensing configuration and the dark-matter properties⁶⁶. This field will soon be revolutionized by upcoming astronomical surveys. The LSST, for instance, is expected to detect more than 8,000 lensed quasars, 13% of which are predicted to be quadruple lenses⁶⁷, which should allow us to probe the subhalo mass function below $10^8 M_\odot$, whereas observations in the optical and near-infrared wavelengths with Euclid and the E-ELT, as well as in radio wavelengths with the Atacama Large Millimeter Array (ALMA) and the global Very-Long-Baseline Interferometry (VLBI) instruments, should allow us to probe the subhalo mass function at high redshift⁶⁸.

Gravitational wave portal

Primordial black holes

The detection of gravitational waves⁶⁹ has opened up new opportunities to explore the physics of dark matter⁷⁰. It has been suggested that the binary black holes whose merger produced the gravitational waves detected by LIGO might be primordial, that is, they might have formed in the very early Universe, before Big Bang nucleosynthesis^{67,71}. The rate of binary black-hole mergers would however be too high if such primordial black holes made up all of the dark matter in the Universe^{72–74}—a possibility that is also disfavoured from a variety of constraints, including the dynamical heating of dwarf galaxies,

distortions of the cosmic microwave background, supernova lensing, and radio and X-ray emission due to the accretion of interstellar gas onto primordial black holes⁷⁵. Although such constraints are becoming stringent, it is important to search for these objects, even if they represent a subdominant component of dark matter. For instance, if we discovered a population of primordial black holes in the Universe, we would know that dark matter is not made of WIMPs, otherwise we should have already detected the annihilation radiation produced by WIMPs around them⁷⁶. A number of observations, such as the identification of black holes lighter than $1M_{\odot}$ or the existence of black holes at a redshift greater than⁷⁷ 40, may in principle provide strong evidence for the existence of primordial black holes.

Constraints on modified gravity

Since a pioneering work on modified Newtonian dynamics published in 1982⁷⁸, numerous attempts have been made (for example, with modified gravity approaches such as the modified gravity model (MOG)⁷⁹ and emergent gravity⁸⁰) to eliminate dark matter by modifying Einstein's theory of general relativity. The success of these efforts, however, remained limited to the rotation curves of galaxies, and it is today clear that the only way that these theories can be reconciled with observations is by mimicking the behaviour of cold dark matter on cosmological scales effectively and very precisely. The coincident observation of gravitational waves and electromagnetic radiation from GW170817⁸¹ has allowed us to set very stringent constraints on the propagation velocity of gravitational waves. The fact that this velocity does not differ from the speed of light by more than one part in 10^{15} severely constrains all modified-gravity theories in which gravitational waves travel on different geodesics with respect to photons and neutrinos^{82–84}. This has in particular allowed us to rule out Bekenstein's tensor–vector–scalar theory⁸⁵.

Black–hole environment

Interestingly, dark matter might manifest itself as a perturbation in the waveform of binary black holes. If dark matter is made of cold and collisionless particles, then their density around black holes will inevitably be higher (possibly much higher) than their average density in the Universe. In particular supermassive black holes at the centre of galaxies might host dark-matter 'spikes'⁸⁶, although dynamical effects, such as mergers with other black holes and interactions with stellar cusps, might disrupt them^{87,88}. Large dark-matter overdensities are possible around intermediate-mass black holes⁸⁹ and around primordial black holes⁹⁰. The presence of dark matter around black holes would modify the dynamics of the merger and induce a potentially detectable dephasing in the waveform⁷⁰. If dark matter is made of ultralight bosons, as in the aforementioned case of fuzzy dark matter, the field 'cloud' that forms around black holes with masses comparable to the Compton wavelength of bosons can be revealed in the gravitational-wave signal from single or binary black holes through direct monochromatic emission, stochastic background or gaps in the black-hole mass–spin Regge plane^{91–93}. Future analyses will allow to further elucidate possible 'environmental' effects due to dark-matter particles and to discriminate among different dark-matter models⁷⁰.

The future

In the quest for dark matter, naturalness has been the guiding principle since the dark-matter problem was established in the early 1980s. Although the absence of evidence for new physics at the LHC does not completely rule out natural theories, we argue that a new era in the search for dark matter has begun, with the new guiding principle being 'no stone left unturned': from fuzzy dark matter (10^{-22} eV) to primordial black holes ($10M_{\odot}$), we should look for dark matter wherever we can. It is important to fully exploit existing experimental facilities—most notably the LHC, whose data might still contain some surprises—and to complete the search for WIMPs with direct-detection experiments until their sensitivity reaches the so-called neutrino floor⁹⁴.

At the same time, we believe that it is essential to diversify the experimental effort and to test the properties of dark matter with gravitational-wave interferometers and upcoming astronomical surveys because they can provide complementary information about the nature of dark matter. New opportunities in extracting such information from data arise from the booming field of machine learning, which is currently transforming many aspects of science and society. Machine-learning methods have been already applied to a variety of dark-matter-related problems, including the identification of WIMPs from particle and astroparticle data^{95,96}, the detection of gravitational lenses⁹⁷, radiation patterns inside quark and gluon jets at the LHC⁹⁸ and real-time gravitational-wave detection⁹⁹. In view of this shift of dark-matter searches towards a more data-driven approach, we believe that it is urgent to fully embrace and, whenever possible, to further develop big-data tools that allow us to organize in a coherent and systematic way the avalanche of data that will become available in particle physics and astronomy in the next decade.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0542-z>

Received: 23 May 2018; Accepted: 6 July 2018;

Published online 3 October 2018.

- Bertone, G. & Hooper, D. A history of dark matter. *Rev. Mod. Phys.* (in the press); preprint at <https://arxiv.org/abs/1605.04909>.
A broad historical perspective on the observational discoveries and the theoretical arguments that led the scientific community to adopt dark matter as an essential part of the standard cosmological model.
- de Swart, J. G., Bertone, G. & van Dongen, J. How dark matter came to matter. *Nat. Astron.* **1**, 0059 (2017).
- Ade, P. A. R. et al. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594**, A13 (2016).
- Bertone, G. et al. *Particle Dark Matter: Observations, Models and Searches* (Cambridge Univ. Press, Cambridge, 2010).
- Hui, L., Ostriker, J. P., Tremaine, S. & Witten, E. Ultralight scalars as cosmological dark matter. *Phys. Rev. D* **95**, 043541 (2017).
- Bird, S. et al. Did LIGO detect dark matter? *Phys. Rev. Lett.* **116**, 201301 (2016).
Shortly after the LIGO detection of gravitational waves, this paper revived the hypothesis that dark matter is made of primordial black holes.
- Clesse, S. & García-Bellido, J. Detecting the gravitational wave background from primordial black hole dark matter. *Phys. Dark Universe* **18**, 105–114 (2017).
- Bertone, G., Hooper, D. & Silk, J. Particle dark matter: evidence, candidates and constraints. *Phys. Rep.* **405**, 279–390 (2005).
- de Gouvêa, A., Hernández, D. & Tait, T. M. P. Criteria for natural hierarchies. *Phys. Rev. D* **89**, 115005 (2014).
- Dine, M. Naturalness under stress. *Annu. Rev. Nucl. Part. Sci.* **65**, 43–62 (2015).
- Bertone, G. The moment of truth for WIMP dark matter. *Nature* **468**, 389–393 (2010).
This 2010 review anticipated that absence of evidence for WIMPs within 5 to 10 years would inevitably lead to the decline of the WIMP paradigm.
- Giudice, G. F. The dawn of the post-naturalness era. Preprint at <https://arxiv.org/abs/1710.07663> (2017).
This article argued that after decades of particle physics research driven by naturalness arguments, we are now witnessing the dawn of the 'post-naturalness' era.
- Athron, P. et al. Global fits of GUT-scale SUSY models with GAMBIT. *Eur. Phys. J. C* **77**, 824 (2017).
- van Beekveld, M., Beenakker, W., Caron, S., Peeters, R. & Ruiz de Austri, R. Supersymmetry with dark matter is still natural. *Phys. Rev. D* **96**, 035015 (2017).
- Ross, G. G., Schmidt-Hoberg, K. & Staub, F. Revisiting fine-tuning in the MSSM. *J. High Energy Phys.* **03**, 021 (2017).
- Goodman, J. et al. Constraints on dark matter from colliders. *Phys. Rev. D* **82**, 116010 (2010).
This paper showed that colliders may detect dark matter even in cases where the expected direct and indirect detection signals are highly suppressed.
- Abdallah, J. et al. Simplified models for dark matter searches at the LHC. *Phys. Dark Universe* **9–10**, 8–23 (2015).
- Beltran, M., Hooper, D., Kolb, E. W., Krusberg, Z. A. C. & Tait, T. M. P. Maverick dark matter at colliders. *J. High Energy Phys.* **09**, 037 (2010).
- Essig, R., Mardon, J. & Volansky, T. Direct detection of sub-GeV dark matter. *Phys. Rev. D* **85**, 076007 (2012).
- Hochberg, Y., Zhao, Y. & Zurek, K. M. Superconducting detectors for superlight dark matter. *Phys. Rev. Lett.* **116**, 011301 (2016).
- Knapen, S., Lin, T. & Zurek, K. M. Light dark matter in superfluid helium: detection with multi-excitation production. *Phys. Rev. D* **95**, 056019 (2017).

22. Hochberg, Y. et al. Detection of sub-MeV dark matter with three-dimensional Dirac materials. *Phys. Rev. D* **97**, 015004 (2018).
23. Essig, R., Schuster, P. & Toro, N. Probing dark forces and light hidden sectors at low-energy e^+e^- colliders. *Phys. Rev. D* **80**, 015003 (2009).
24. Silverwood, H., Weniger, C., Scott, P. & Bertone, G. A realistic assessment of the CTA sensitivity to dark matter annihilation. *J. Cosmol. Astropart. Phys.* **1503**, 055 (2015).
25. Acharya, B. S. et al. Science with the Cherenkov Telescope Array. Preprint at <https://arxiv.org/abs/1709.07997> (2017).
26. Abbott, L. F. & Sikivie, P. A cosmological bound on the invisible axion. *Phys. Lett. B* **120**, 133–136 (1983).
27. Du, N. et al. A search for invisible axion dark matter with the Axion Dark Matter Experiment. *Phys. Rev. Lett.* **120**, 151301 (2018).
28. Kahn, Y., Safdi, B. R. & Thaler, J. Broadband and resonant approaches to axion dark matter detection. *Phys. Rev. Lett.* **117**, 141801 (2016).
29. Graham, P. W., Irastorza, I. G., Lamoreaux, S. K., Lindner, A. & van Bibber, K. A. Experimental searches for the axion and axion-like particles. *Annu. Rev. Nucl. Part. Sci.* **65**, 485–514 (2015).
- A comprehensive review of present and upcoming experimental searches for axions and axion-like particles.**
30. Caldwell, A. et al. Dielectric haloscopes: a new way to detect axion dark matter. *Phys. Rev. Lett.* **118**, 091801 (2017).
31. Shi, X.-D. & Fuller, G. M. A new dark matter candidate: nonthermal sterile neutrinos. *Phys. Rev. Lett.* **82**, 2832–2835 (1999).
32. Laine, M. & Shaposhnikov, M. Sterile neutrino dark matter as a consequence of nuMSM-induced lepton asymmetry. *J. Cosmol. Astropart. Phys.* **0806**, 031 (2008).
33. Boyarsky, A., Ruchayskiy, O. & Shaposhnikov, M. The role of sterile neutrinos in cosmology and astrophysics. *Annu. Rev. Nucl. Part. Sci.* **59**, 191–214 (2009).
34. Drewes, M. et al. A white paper on keV sterile neutrino dark matter. *J. Cosmol. Astropart. Phys.* **1701**, 025 (2017).
35. Bulbul, E. et al. Detection of an unidentified emission line in the stacked X-ray spectrum of galaxy clusters. *Astrophys. J.* **789**, 13 (2014).
36. Jeltema, T. E. & Profumo, S. Discovery of a 3.5 keV line in the Galactic Centre and a critical look at the origin of the line across astronomical targets. *Mon. Not. R. Astron. Soc.* **450**, 2143–2152 (2015).
37. Abazajian, K. N. Sterile neutrinos in cosmology. *Phys. Rep.* **711–712**, 1–28 (2017).
- A comprehensive review of the astroparticle and cosmological aspects of sterile neutrinos.**
38. Kolb, E. W., Chung, D. J. H. & Riotto, A. WIMPzillas! *AIP Conf. Proc.* **484**, 91–105 (1999).
39. Berezhiani, L. & Khoury, J. Theory of dark matter superfluidity. *Phys. Rev. D* **92**, 103510 (2015).
40. Kuhnel, F., Starkman, G. D., Freese, K. & Matas, A. Primordial black-hole and macroscopic dark-matter constraints with LISA. Preprint at <https://arxiv.org/abs/1705.10361> (2017).
41. Buckley, M. R. & Peter, A. H. G. Gravitational probes of dark matter physics. Preprint at <https://arxiv.org/abs/1712.06615> (2017).
- A review of the expected impact of future astrophysical measurements on our understanding of dark matter.**
42. Riess, A. G. et al. New parallaxes of Galactic Cepheids from spatially scanning the Hubble Space Telescope: implications for the Hubble constant. *Astrophys. J.* **855**, 136 (2018).
43. Frenk, C. S. & White, S. D. M. Dark matter and cosmic structure. *Ann. Phys.* **524**, 507–534 (2012).
44. Spergel, D. N. & Steinhardt, P. J. Observational evidence for self-interacting cold dark matter. *Phys. Rev. Lett.* **84**, 3760–3763 (2000).
45. Tulin, S. & Yu, H.-B. Dark matter self-interactions and small scale structure. *Phys. Rep.* **730**, 1–57 (2018).
46. Brinckmann, T., Zavala, J., Rapetti, D., Hansen, S. H. & Vogelsberger, M. The structure and assembly history of cluster-sized haloes in self-interacting dark matter. *Mon. Not. R. Astron. Soc.* **474**, 746–759 (2018).
47. Robertson, A. et al. The diverse density profiles of galaxy clusters with self-interacting dark matter plus baryons. *Mon. Not. R. Astron. Soc.* **476**, L20 (2018).
48. Kaplinghat, M., Keeley, R. E., Linden, T. & Yu, H.-B. Tying dark matter to baryons with selfinteractions. *Phys. Rev. Lett.* **113**, 021302 (2014).
49. Harvey, D., Massey, R., Kitching, T., Taylor, A. & Tittley, E. The non-gravitational interactions of dark matter in colliding galaxy clusters. *Science* **347**, 1462–1465 (2015).
50. Robertson, A., Massey, R. & Eke, V. Cosmic particle colliders: simulations of self-interacting dark matter with anisotropic scattering. *Mon. Not. R. Astron. Soc.* **467**, 4719–4730 (2017).
51. Randall, S. W., Markevitch, M., Clowe, D., Gonzalez, A. H. & Bradac, M. Constraints on the self-interaction cross-section of dark matter from numerical simulations of the merging galaxy cluster 1E 0657–56. *Astrophys. J.* **679**, 1173–1180 (2008).
52. Harvey, D., Courbin, F., Kneib, J. P. & McCarthy, I. G. A detection of wobbling brightest cluster galaxies within massive galaxy clusters. *Mon. Not. R. Astron. Soc.* **472**, 1972–1980 (2017).
53. Narayanan, V. K., Spergel, D. N., Dave, R. & Ma, C.-P. Constraints on the mass of warm dark matter particles and the shape of the linear power spectrum from the Ly α forest. *Astrophys. J.* **543**, L103–L106 (2000).
54. Iršič, V. et al. New Constraints on the free-streaming of warm dark matter from intermediate and small scale Lyman- α forest data. *Phys. Rev. D* **96**, 023522 (2017).
55. Iršič, V., Viel, M., Haehnelt, M. G., Bolton, J. S. & Becker, G. D. First constraints on fuzzy dark matter from Lyman- α forest data and hydrodynamical simulations. *Phys. Rev. Lett.* **119**, 031302 (2017).
56. Yoon, J. H., Johnston, K. V. & Hogg, D. W. Clumpy streams from clumpy halos: detecting missing satellites with cold stellar structures. *Astrophys. J.* **731**, 58 (2011).
57. Carlberg, R. G. Dark matter sub-halo counts via star stream crossings. *Astrophys. J.* **748**, 20 (2012).
58. Bovy, J., Erkal, D. & Sanders, J. L. Linear perturbation theory for tidal streams and the small-scale CDM power spectrum. *Mon. Not. R. Astron. Soc.* **466**, 628–668 (2017).
59. Erkal, D. & Belokurov, V. Properties of dark subhaloes from gaps in tidal streams. *Mon. Not. R. Astron. Soc.* **454**, 3542–3558 (2015).
60. Banik, N., Bertone, G., Bovy, J. & Bozorgnia, N. Probing the nature of dark matter particles with stellar streams. *J. Cosmol. Astropart. Phys.* **7**, 061 (2018).
61. Mao, S. & Schneider, P. Evidence for substructure in lens galaxies? *Mon. Not. R. Astron. Soc.* **295**, 587 (1998).
62. Metcalf, R. B. & Madau, P. Compound gravitational lensing as a probe of dark matter substructure within galaxy halos. *Astrophys. J.* **563**, 9–20 (2001).
63. Dalal, N. & Kochanek, C. S. Direct detection of cold dark matter substructure. *Astrophys. J.* **572**, 25–33 (2002).
64. Gilman, D., Birrer, S., Treu, T. & Keeton, C. R. Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. *Mon. Not. R. Astron. Soc.* <https://doi.org/10.1093/mnras/sty2261> (2018).
65. Vegetti, S. & Koopmans, L. V. E. Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in galaxies. *Mon. Not. R. Astron. Soc.* **392**, 945 (2009).
66. Despali, G., Vegetti, S., White, S. D. M., Giocoli, C. & van den Bosch, F. C. Modelling the line-of-sight contribution in substructure lensing. *Mon. Not. R. Astron. Soc.* **475**, 5424–5442 (2018).
67. Oguri, M. & Marshall, P. J. Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. *Mon. Not. R. Astron. Soc.* **405**, 2579–2593 (2010).
68. Daylan, T., Cyr-Racine, F.-Y., Diaz Rivero, A., Dvorkin, C. & Finkbeiner, D. P. Probing the small-scale structure in strongly lensed systems via transdimensional inference. *Astrophys. J.* **854**, 141 (2018).
69. Abbott, B. P. et al. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.* **116**, 061102 (2016).
70. Barack, L. et al. Black holes, gravitational waves and fundamental physics: a roadmap. Preprint at <https://arxiv.org/abs/1806.05195> (2018).
- This article contains a discussion about the role of gravitational waves in the search for dark matter.**
71. Carr, B., Kuhnel, F. & Sandstad, M. Primordial black holes as dark matter. *Phys. Rev. D* **94**, 083504 (2016).
72. Sasaki, M., Suyama, T., Tanaka, T. & Yokoyama, S. Primordial black hole scenario for the gravitational-wave event GW150914. *Phys. Rev. Lett.* **117**, 061101 (2016).
73. Ali-Haïmoud, Y., Kovetz, E. D. & Kamionkowski, M. Merger rate of primordial black-hole binaries. *Phys. Rev. D* **96**, 123523 (2017).
74. Kavanagh, B. J., Gaggero, D. & Bertone, G. Merger rate of a subdominant population of primordial black holes. *Phys. Rev. D* **98**, 023536 (2018).
75. Gaggero, D. et al. Searching for primordial black holes in the radio and X-ray sky. *Phys. Rev. Lett.* **118**, 241101 (2017).
76. Lacki, B. C. & Beacom, J. F. Primordial black holes as dark matter: almost all or almost nothing. *Astrophys. J.* **720**, L67–L71 (2010).
77. Koushiappas, S. M. & Loeb, A. Maximum redshift of gravitational wave merger events. *Phys. Rev. Lett.* **119**, 221104 (2017).
78. Milgrom, M. A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *Astrophys. J.* **270**, 365–370 (1983).
79. Moffat, J. W. Scalar-tensor-vector gravity theory. *J. Cosmol. Astropart. Phys.* **0603**, 004 (2006).
80. Verlinde, E. P. Emergent gravity and the dark Universe. *SciPost Phys.* **2**, 016 (2017).
81. Abbott, B. et al. GW170817: observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.* **119**, 161101 (2017).
82. Boran, S., Desai, S., Kahya, E. O. & Woodard, R. P. GW170817 falsifies dark matter emulators. *Phys. Rev. D* **97**, 041501 (2018).
83. Sakstein, J. & Jain, B. Implications of the neutron star merger GW170817 for cosmological scalar-tensor theories. *Phys. Rev. Lett.* **119**, 251303 (2017).
84. Wang, H. et al. The GW170817/GRB 170817A/AT 2017gfo association: some implications for physics and astrophysics. *Astrophys. J.* **851**, L18 (2017).
85. Bekenstein, J. D. Relativistic gravitation theory for the MOND paradigm. *Phys. Rev. D* **70**, 083509 (2004); erratum **71**, 069901 (2005).
86. Gondolo, P. & Silk, J. Dark matter annihilation at the galactic center. *Phys. Rev. Lett.* **83**, 1719–1722 (1999).
87. Merritt, D., Milosavljevic, M., Verde, L. & Jimenez, R. Dark matter spikes and annihilation radiation from the galactic center. *Phys. Rev. Lett.* **88**, 191301 (2002).
88. Bertone, G. & Merritt, D. Time-dependent models for dark matter at the Galactic Center. *Phys. Rev. D* **72**, 103502 (2005).
89. Bertone, G., Zentner, A. R. & Silk, J. A new signature of dark matter annihilations: gamma-rays from intermediate-mass black holes. *Phys. Rev. D* **72**, 103517 (2005).
90. Ricotti, M., Ostriker, J. P. & Mack, K. J. Effect of primordial black holes on the cosmic microwave background and cosmological parameter estimates. *Astrophys. J.* **680**, 829–845 (2008).

91. Brito, R. et al. Gravitational wave searches for ultralight bosons with LIGO and LISA. *Phys. Rev. D* **96**, 064050 (2017).
 92. Arvanitaki, A., Baryakhtar, M., Dimopoulos, S., Dubovsky, S. & Lasenby, R. Black hole mergers and the QCD axion at advanced LIGO. *Phys. Rev. D* **95**, 043001 (2017).
 93. Baumann, D., Chia, H. S. & Porto, R. A. Probing ultralight bosons with binary black holes. Preprint at <https://arxiv.org/abs/1804.03208> (2018).
 94. Billard, J., Strigari, L. & Figueroa-Feliciano, E. Implication of neutrino backgrounds on the reach of next generation dark matter direct detection experiments. *Phys. Rev. D* **89**, 023524 (2014).
 95. Bertone, G. et al. Identifying WIMP dark matter from particle and astroparticle data. *J. Cosmol. Astropart. Phys.* **1803**, 026 (2018).
 96. Caron, S., Kim, J. S., Rolbiecki, K., Ruiz de Austri, R. & Stienen, B. The BSM-AI project: SUSYAI-generalizing LHC limits on supersymmetry with machine learning. *Eur. Phys. J. C* **77**, 257 (2017).
 97. Hezaveh, Y. D., Perreault Levasseur, L. & Marshall, P. J. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature* **548**, 555–557 (2017).
- An interesting example of the application of machine-learning methods to dark matter studies.**
98. Larkoski, A. J., Mout, I. & Nachman, B. Jet substructure at the Large Hadron Collider: a review of recent advances in theory and machine learning. Preprint at <https://arxiv.org/abs/1709.04464> (2017).
 99. George, D. & Huerta, E. A. Deep learning for real-time gravitational wave detection and parameter estimation: results with advanced LIGO data. *Phys. Lett. B* **778**, 64–70 (2018).

Acknowledgements We thank V. Cardoso, D. Gaggero, D. Harvey, D. Hooper, B. Kavanagh, S. Vegetti and M. Viel for comments on the initial version of this manuscript. The work of T.M.P.T. is supported in part by NSF grant PHY-1316792.

Reviewer information *Nature* thanks M. Kamionkowski and R. Massey for their contribution to the peer review of this work.

Author contributions G.B. conceived the idea of the review. G.B. and T.M.P.T. contributed equally to the writing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.B. or T.M.P.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Plant functional trait change across a warming tundra biome

A list of authors and their affiliations appears online.

The tundra is warming more rapidly than any other biome on Earth, and the potential ramifications are far-reaching because of global feedback effects between vegetation and climate. A better understanding of how environmental factors shape plant structure and function is crucial for predicting the consequences of environmental change for ecosystem functioning. Here we explore the biome-wide relationships between temperature, moisture and seven key plant functional traits both across space and over three decades of warming at 117 tundra locations. Spatial temperature–trait relationships were generally strong but soil moisture had a marked influence on the strength and direction of these relationships, highlighting the potentially important influence of changes in water availability on future trait shifts in tundra plant communities. Community height increased with warming across all sites over the past three decades, but other traits lagged far behind predicted rates of change. Our findings highlight the challenge of using space-for-time substitution to predict the functional consequences of future warming and suggest that functions that are tied closely to plant height will experience the most rapid change. They also reveal the strength with which environmental factors shape biotic communities at the coldest extremes of the planet and will help to improve projections of functional changes in tundra ecosystems with climate warming.

Rapid climate warming in Arctic and alpine regions is driving changes in the structure and composition of tundra ecosystems^{1,2}, with potentially global consequences. Up to 50% of the world's belowground carbon stocks are contained in permafrost soils³, and tundra regions are expected to contribute the majority of warming-induced soil carbon loss over the next century⁴. Plant traits strongly affect carbon cycling and the energy balance of the ecosystem, which can in turn influence regional and global climates^{5–7}. Traits related to the resource economics spectrum⁸, such as specific leaf area (SLA), leaf nitrogen content and leaf dry matter content (LDMC), affect primary productivity, litter decomposability, soil carbon storage and nutrient cycling^{5,6,9,10}, while size-related traits, such as leaf area and plant height, influence aboveground carbon storage, albedo (that is, surface reflectance) and hydrology^{11–13} (Extended Data Table 1). Quantifying the link between the environment and plant functional traits is therefore important to understanding the consequences of climate change, but such studies rarely extend into the tundra^{14–16}. Thus, the full extent of the relationship between climate and plant traits in the coldest ecosystems on Earth has yet to be assessed, and the consequences of climate warming for functional change in the tundra remain largely unknown.

Here we quantify the biome-wide relationships between temperature, soil moisture and key traits that represent the foundation of plant form and function¹⁷, using a dataset of more than 56,000 tundra plant trait observations (Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). We examine five continuously distributed traits related to plant size (adult plant height and leaf area) and to resource economy (SLA, leaf nitrogen content and LDMC), as well as two categorical traits related to community-level structure (woodiness) and leaf phenology and lifespan (evergreenness). Intraspecific trait variability is thought to be especially important in regions where diversity is low or where species have wide geographical ranges¹⁸, as in the tundra. Thus, we analyse two underlying components of biogeographical patterns in the five continuous traits: intraspecific variability (phenotypic plasticity or genetic differences among populations) and community-level variability (species turnover or shifts in the abundances of species across space). We first investigated how plant traits vary with temperature

and soil moisture across the tundra biome. We then quantified the relative influence of intraspecific trait variation (ITV) versus community-level trait variation (estimated as community-weighted trait means (CWM)) for spatial temperature–trait relationships. Finally, we investigated whether spatial temperature–trait relationships are explained by among-site differences in species abundance or species turnover (presence or absence).

A major incentive for quantifying spatial temperature–trait relationships is to provide an empirical basis for predicting the potential consequences of future warming^{19–21}. Thus, we also estimate realized rates of community-level trait change over time using nearly three decades of vegetation survey data at 117 tundra sites (Fig. 1a and Supplementary Table 2). Focusing on interspecific trait variation, we investigated how changes in community traits over three decades of ambient warming compare to predictions from spatial temperature–trait relationships. We expect greater temporal trait change when spatial temperature–trait relationships are (a) strong, (b) unlimited by moisture availability and (c) due primarily to abundance shifts instead of species turnover, given that species turnover over time depends on immigration and is likely to be slow²². Finally, because total realized trait change in continuous traits consists of both community-level variation and ITV, we estimated the potential contribution of ITV to overall trait change (CWM + ITV) using the modelled intraspecific temperature–trait relationships described above (see Methods and Extended Data Fig. 1b). For all analyses, we used a generalizable Bayesian modelling approach, which allowed us to account for the hierarchical spatial, temporal and taxonomic structure of the data as well as multiple sources of uncertainty.

Environment–trait relationships across the tundra biome

We found strong spatial associations between temperature and community height, SLA and LDMC (Fig. 2a, Extended Data Fig. 2 and Supplementary Table 3) across the 117 survey sites. Both height and SLA increased with summer temperature, but the temperature–trait relationship for SLA was much stronger at wetter than at drier sites. LDMC was negatively related to temperature, and

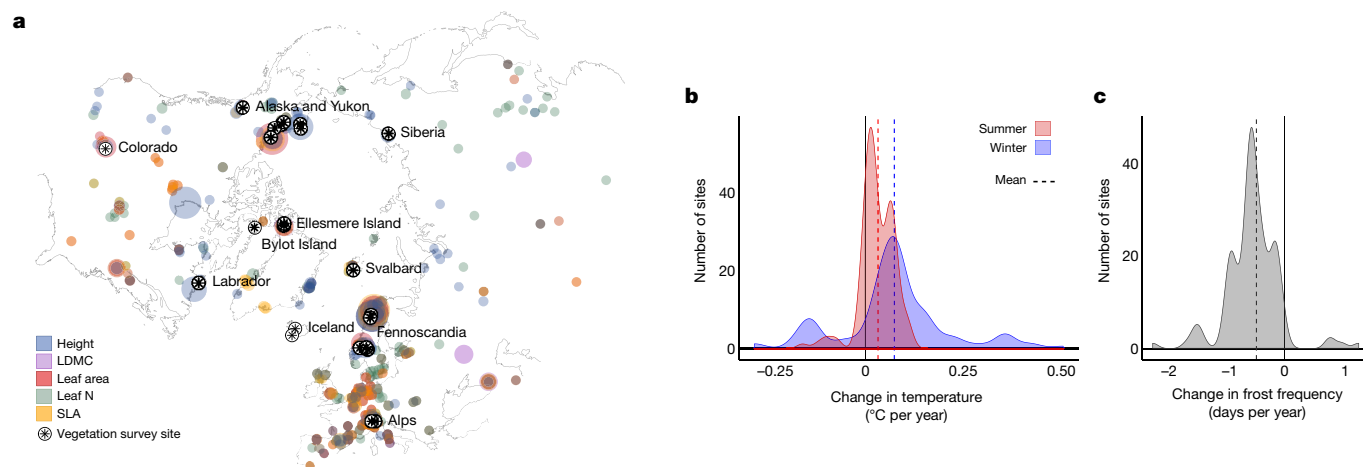


Fig. 1 | Geographical distribution of trait and vegetation survey data and climatic change over the study period. **a**, Map of all 56,048 tundra trait records and 117 vegetation survey sites. **b**, **c**, Climatic change across the period of monitoring at the 117 vegetation survey sites, represented as mean winter (coldest quarter) and summer (warmest quarter) temperature (**b**) and frost day frequency (**c**). The size of the coloured points on the map indicates the relative quantity of trait measurements (larger circles indicate more measurements of that trait at a given location) and the colour indicates which trait was measured. The black stars indicate the vegetation

more strongly so at wetter than drier sites. Community woodiness decreased with temperature, but the ratio of evergreen to deciduous woody species increased with temperature, particularly at drier sites (Extended Data Fig. 3). These spatial temperature–trait relationships indicate that long-term climate warming should cause pronounced shifts towards communities of taller plants with more resource-acquisitive leaves (high SLA and low LDMC), particularly where soil moisture is high.

Our results reveal a substantial moderating influence of soil moisture on community traits across spatial temperature gradients^{2,23}. Both leaf area and leaf nitrogen content decreased with warmer temperatures in dry sites but increased with warmer temperatures in wet sites (Fig. 2a and Supplementary Table 4). Soil moisture was important for explaining spatial variation in all seven investigated traits, even when temperature alone was not (for example, leaf area; Fig. 2a and Extended Data Fig. 2), potentially reflecting physiological constraints that are related to heat exchange or frost tolerance when water availability is low²⁴. Thus, future warming-driven changes in traits and associated ecosystem functions (for example, decomposability) will probably depend on current soil moisture conditions at a site²³. Furthermore, future changes in water availability (for example, because of changes in precipitation, snow melt timing, permafrost and hydrology²⁵) could cause substantial shifts in these traits and their associated functions, irrespective of warming.

We found consistent intraspecific temperature–trait relationships for all five continuous traits (Fig. 2b and Supplementary Table 5). Intraspecific plant height and leaf area showed strong positive relationships with summer temperature (that is, individuals were taller and had larger leaves in warmer locations), whereas intraspecific LDMC, leaf nitrogen content and SLA were related to winter but not summer temperature (Extended Data Fig. 2). The differences in responses of ITV to summer versus winter temperatures may indicate that size-related traits better reflect summer growth potential, whereas resource-economics traits reflect tolerance to cold-stress. These results, although correlative, indicate that trait variation expressed at the individual or population level is related to the growing environment and that warming will probably lead to substantial intraspecific change in many traits. Thus, the potential for trait change over time is underestimated by using species-level trait means alone. Future work is needed to disentangle

survey sites used in the community trait analyses (most stars represent multiple sites). Trait data were included for all species that occurred in at least one tundra vegetation survey site; thus, although not all species are unique to the tundra, all do occur in the tundra. Temperature change and frost frequency change were estimated for the interval over which sampling was conducted at each site plus the preceding four years, to best reflect the time window over which tundra plant communities respond to temperature change^{20,29}.

the role of plasticity and genetic differentiation in explaining the observed intraspecific temperature–trait relationships²⁶, as this will also influence the rate of future trait change²⁷. Trait measurements collected over time and under novel (experimental) conditions, as yet unavailable, would enable more accurate predictions of future intraspecific trait change.

Partitioning the underlying causes of community temperature–trait relationships revealed that species turnover explained most of the variation in traits across space (Fig. 2c), suggesting that dispersal and immigration processes will primarily govern the rate of ecosystem responses to warming. Shifts in the abundances of species and ITV accounted for a relatively small part of the overall temperature–trait relationship across space (Fig. 2c). Furthermore, the local trait pool in the coldest tundra sites (mean summer temperature $<3^{\circ}\text{C}$) is constrained relative to the tundra as a whole for many traits (Extended Data Fig. 4). Together, these results indicate that the magnitude of warming-induced community trait shifts will be limited without the arrival of novel species from warmer environments.

Change in community traits over time

Plant height was the only trait for which the CWM changed over the 27 years of monitoring; it increased rapidly at nearly every survey site (Fig. 3a, b, Extended Data Fig. 3 and Supplementary Table 6). Interannual variation in community height was sensitive to summer temperature (Fig. 3c, Extended Data Fig. 2 and Supplementary Table 7), indicating that increases in community height are responding to warming. However, neither the total rate of temperature change nor soil moisture predicted the total rate of CWM change in any trait (Extended Data Fig. 5 and Supplementary Table 8). Incorporating potential ITV doubled the average estimate of plant height change over time (Figs. 3a, 4a, dashed lines). Because spatial patterns in ITV can be due to both phenotypic plasticity and genetic differences among populations, this is likely to be a maximum estimate of the ITV contribution to trait change (for example, if intraspecific temperature–trait relationships are due entirely to phenotypic plasticity). The observed increase in community height is consistent with previous findings of increasing vegetation height in response to experimental warming at a subset of these sites²⁸ and with studies showing increased shrub growth over time¹¹.

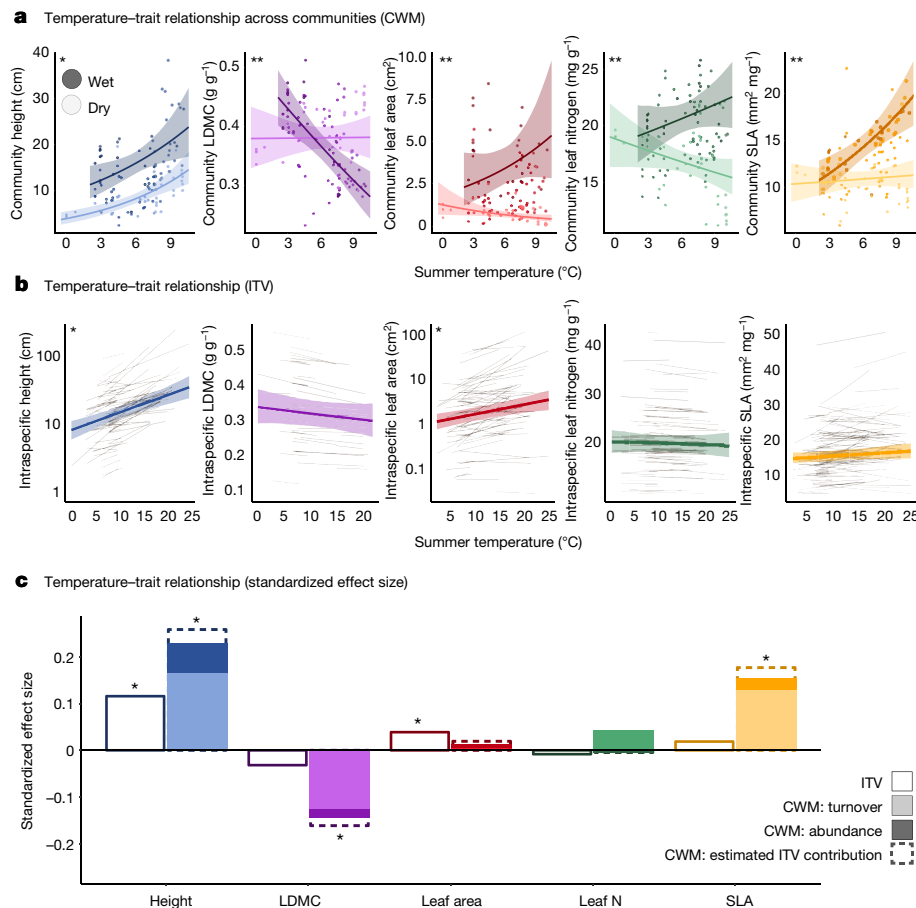


Fig. 2 | Strong spatial relationships in traits across temperature and soil moisture gradients are primarily explained by species turnover. **a**, Spatial relationship between community-level (CWM) functional traits, mean summer (warmest quarter) temperature and soil moisture ($n = 1,520$ plots within 117 sites within 72 regions). **b**, Spatial relationship between summer temperature and ITV (note the log scale for height and leaf area). **c**, Standardized effect sizes were estimated for all temperature–trait relationships both across communities (CWM; solid bars) and within species (ITV; open bars with solid outlines). Effect sizes for CWM temperature–trait relationships were further partitioned into the proportion of the effect driven solely by species turnover (light bars) and abundance shifts (dark bars) over space. Dashed lines indicate the estimated additional contribution of ITV to the total temperature–trait relationship (CWM + ITV). The contribution of ITV is estimated from

the spatial temperature–trait relationships modelled in **b**. Soil moisture in **a** was modelled as continuous but is shown predicted only at low and high values to improve visualization. Transparent ribbons in **a** and **b** indicate 95% credible intervals for model mean predictions. Grey lines in **b** represent intraspecific temperature–trait relationships for each species (height, $n = 80$ species; LDMC, $n = 43$; leaf area, $n = 85$; leaf nitrogen content (leaf N), $n = 85$; SLA, $n = 108$; the number of observations per trait is shown in Supplementary Table 1). In all panels, asterisks indicate that the 95% credible interval on the slope of the temperature–trait relationship did not overlap zero. In **a**, two asterisks indicate that the temperature \times soil moisture interaction term did not overlap zero. Winter temperature–trait relationships are shown in Extended Data Fig. 2. Community woodiness and evergreenness are shown in Extended Data Fig. 3.

Increasing community height over time was mostly attributable to species turnover (rather than shifts in abundance of the resident species; Fig. 3b) and was driven by the immigration of taller species rather than the loss of shorter ones (Extended Data Fig. 6 and Supplementary Table 9). This turnover could reflect the movement of tall species upward in latitude and elevation or from local species pools in nearby warmer microclimates. The magnitude of temporal change was comparable to the change predicted based on the spatial temperature–trait relationship (Fig. 4a, solid lines), indicating that temporal change in plant height is not currently limited by immigration rates. The importance of immigration in explaining changes in community height is surprising given the relatively short study duration and long lifespan of tundra plants, but is nonetheless consistent with a previous finding of shifts towards warm-associated species in tundra plant communities^{20,29}. If the observed rate of trait change continues (for example, if immigration were unlimited), community height (excluding potential change due to ITV) could increase by 20–60% by the end of the century, depending on carbon emission, warming and water availability scenarios (Extended Data Fig. 7).

Consequences and implications

Recent (observed) and future (predicted) changes in plant traits, particularly height, are likely to have important implications for ecosystem functions and feedback effects involving soil temperature^{30,31}, decomposition^{5,10} and carbon cycling³², as the potential for soil carbon loss is particularly great in high-latitude regions⁴. For example, increasing plant height could offset warming-driven carbon loss through increased carbon storage due to woody litter production⁵ or through reduced decomposition owing to lower summer soil temperatures caused by shading^{3,30,32} (negative feedback effects). Positive feedback effects are also possible if branches or leaves above the snowpack reduce albedo^{11,12} or increase snow accumulation, leading to warmer soil temperatures in winter and increased decomposition rates^{3,11}. The balance of these feedback systems—and thus the net effect of trait change on carbon cycling—may depend on the interaction between warming and changes in snow distribution³³ and water availability³⁴, which remain mostly unknown for the tundra biome.

The lack of an observed temporal trend in SLA and LDMC, despite strong temperature–trait relationships over space, highlights the limitations of using space-for-time substitution for predicting

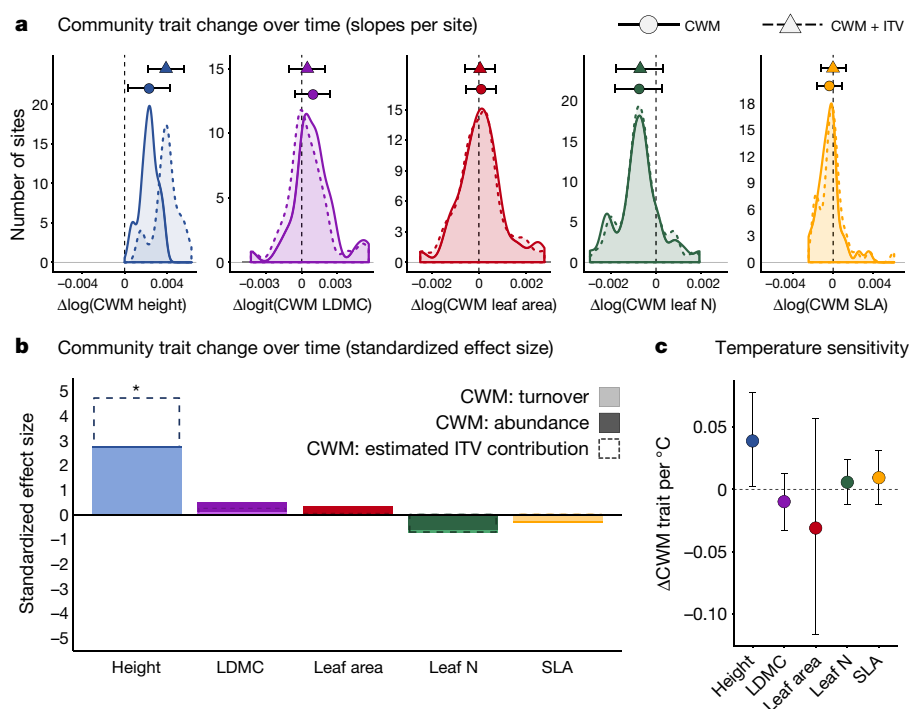


Fig. 3 | A tundra-wide increase in community height over time is related to warming. **a**, Observed community trait change per year (transformed units). Solid lines indicate the distribution of CWM model slopes (trait change per site) whereas dashed lines indicate change in CWM plus potential intraspecific change modelled from spatial temperature–trait relationships (CWM + ITV). Circles (CWM) or triangles (CWM + ITV) and error bars indicate the mean and 95% credible interval for the overall rate of trait change across all sites ($n = 4,575$ plot-years within 117 sites within 38 regions). The vertical black dashed line indicates 0 (no change over time). **b**, Standardized effect sizes for CWM change over time were further partitioned into the proportion of the effect driven solely by species turnover (light bars) or shifts in abundance of resident species (dark bars) over time. Dashed lines indicate the estimated

additional contribution of ITV to total trait change over time (CWM + ITV). Asterisks indicate that the 95% credible interval on the mean hyperparameter for CWM trait change over time did not overlap zero. **c**, Temperature sensitivity of each trait (that is, correspondence between interannual variation in CWM trait values and interannual variation in summer temperature). Temperatures associated with each survey year were estimated as five-year means (temperature of the survey year and four preceding years), because this interval has been shown to be most relevant to vegetation change in tundra²⁰ and alpine²⁹ plant communities. Circles represent the mean temperature sensitivity across all 117 sites, error bars are 95% credible intervals on the mean. Changes in community woodiness and evergreenness are shown in Extended Data Fig. 3.

short-term ecological change. This disconnect could reflect the influence of unmeasured changes in water availability (for example, owing to local-scale variation in the timing of snowmelt or hydrology) that counter or overwhelm the effect of static soil moisture estimates. For example, we would not expect substantial changes in traits demonstrating a spatial temperature \times moisture interaction (LDMC, leaf area, leaf nitrogen content and SLA), even in wet sites, if warming also leads to drier soils. Plant height was the only continuous trait for which a temperature \times moisture interaction was not important, and was predicted to increase across all areas of the tundra regardless of recent soil moisture trends (Fig. 4c, d). Spatiotemporal disconnects could also reflect dispersal limitation of potential immigrants (for example, with low LDMC and high SLA) or establishment failure due to novel biotic³⁵ or abiotic³⁶ conditions other than temperature to which immigrants are maladapted^{22,36}. Furthermore, community responses to climate warming could be constrained by soil properties (for example, organic matter and mineralization) that themselves respond slowly to warming²⁰.

The patterns in functional traits described here reveal the extent to which environmental factors shape biotic communities in the tundra. Strong temperature- and moisture-related spatial gradients in traits related to competitive ability (for example, height) and resource capture and retention (for example, leaf nitrogen and SLA) reflect trade-offs in plant ecological strategy^{9,37} from benign (warm, wet) to extreme (cold, dry) conditions. Community-level trait syndromes, as reflected in ordination axes, are also strongly related to both temperature and moisture, suggesting that environmental drivers structure not only individual traits but also trait combinations—and thus lead to a limited number

of successful functional strategies in some environments (for example, woody, low-SLA and low-leaf nitrogen communities in warm, dry sites; Extended Data Fig. 8). Thus, warming may lead to a community-level shift towards more acquisitive plant strategies³⁷ in wet tundra sites, but towards more conservative strategies in drier sites as moisture becomes more limiting.

Earth system models are increasingly moving to incorporate relationships between traits and the environment, as this can substantially improve estimates of ecosystem change^{38–40}. Our results inform these projections of future tundra functional change³⁸ by explicitly quantifying the link between temperature, moisture and key functional traits across the biome. In particular, our study highlights the importance of accounting for future changes in water availability, as this will probably influence both the magnitude and direction of change for many traits. In addition, we demonstrate that spatial trait–environment relationships are driven largely by species turnover, suggesting that modelling efforts must account for rates of species immigration when predicting the speed of future functional shifts. The failure of many traits (for example, SLA) to match expected rates of change suggests that space-for-time substitution alone may inaccurately represent near-term ecosystem change. Nevertheless, the ubiquitous increase in community plant height reveals that functional change is already occurring in tundra ecosystems.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0563-7>.

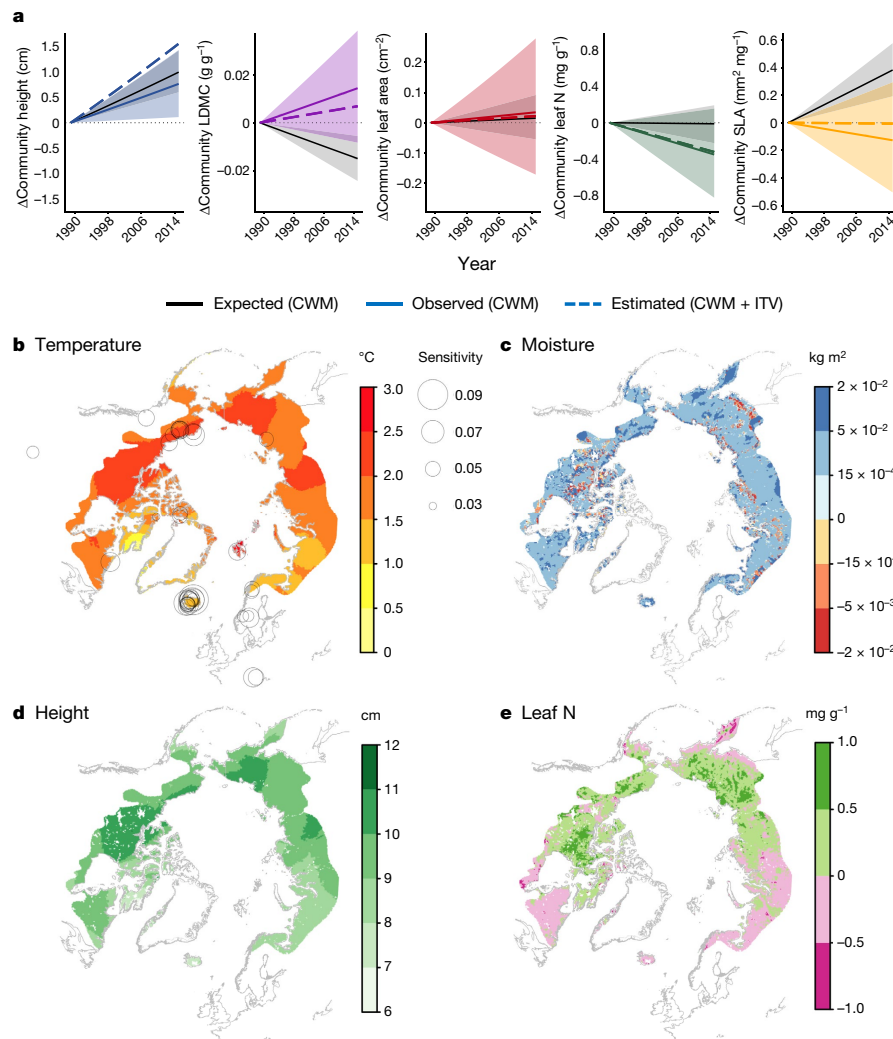


Fig. 4 | Community height increases in line with space-for-time predictions but other traits lag. **a**, Observed community (CWM) trait change over time (coloured lines) across all 117 sites versus expected CWM change over the duration of vegetation monitoring (1989–2015) based on the spatial temperature–trait (CWM) relationship and the average rate of recent summer warming across all sites (solid black lines). Coloured dashed lines indicate the estimated total change over time if predicted intraspecific trait variability is also included (CWM + ITV). Values on the y axis represent the magnitude of change relative to 0 (that is, trait anomaly), with 0 representing the trait value at t_0 . **b**, **c**, Total recent temperature change (**b**) and soil moisture change (**c**) across the Arctic tundra (1979–2016). Temperature change estimates are derived from gridded temperature data from the Climate Research Unit (CRU), estimates of changes in soil moisture are derived from downscaled

European Centre for Medium-Range Weather Forecasts Re-Analysis (ERA-Interim) soil moisture data. Circles in **b** represent the sensitivity (cm per °C) of CWM plant height to summer temperature at each site (see Fig. 3c). Areas of high temperature sensitivity are expected to experience the greatest increases in height with warming. **d**, **e**, Spatial trait–temperature–moisture relationships (Fig. 2a) were used to predict total changes in height (**d**) and leaf nitrogen content (**e**) over the entire 1979–2016 period based on concurrent changes in temperature and soil moisture. Note that **d** and **e** reflect the magnitude of expected change between 1979 and 2016, not observed trait change. See Methods for details on estimates of the change in temperature and soil moisture. The outline of Arctic areas is based on the Circumpolar Arctic Vegetation Map (<http://www.geobotany.uaf.edu/cavm>).

Received: 15 September 2017; Accepted: 8 August 2018;
Published online 26 September 2018.

- Post, E. et al. Ecological dynamics across the Arctic associated with recent climate change. *Science* **325**, 1355–1358 (2009).
- Elmendorf, S. C. et al. Plot-scale evidence of tundra vegetation change and links to recent summer warming. *Nat. Clim. Change* **2**, 453–457 (2012).
- Sistla, S. A. et al. Long-term warming restructures Arctic tundra without changing net soil carbon storage. *Nature* **497**, 615–618 (2013).
- Crowther, T. W. et al. Quantifying global soil carbon losses in response to warming. *Nature* **540**, 104–108 (2016).
- Cornelissen, J. H. C. et al. Global negative vegetation feedback to climate warming responses of leaf litter decomposition rates in cold biomes. *Ecol. Lett.* **10**, 619–627 (2007).
- Lavorel, S. & Garnier, E. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Funct. Ecol.* **16**, 545–556 (2002).

- Pearson, R. G. et al. Shifts in Arctic vegetation and associated feedbacks under climate change. *Nat. Clim. Change* **3**, 673–677 (2013).
- Wright, I. J. et al. The worldwide leaf economics spectrum. *Nature* **428**, 821–827 (2004).
- Díaz, S. et al. The plant traits that drive ecosystems: evidence from three continents. *J. Veg. Sci.* **15**, 295–304 (2004).
- Cornwell, W. K. et al. Plant species traits are the predominant control on litter decomposition rates within biomes worldwide. *Ecol. Lett.* **11**, 1065–1071 (2008).
- Myers-Smith, I. H. et al. Shrub expansion in tundra ecosystems: dynamics, impacts and research priorities. *Environ. Res. Lett.* **6**, 045509 (2011).
- Sturm, M. & Douglas, T. Changing snow and shrub conditions affect albedo with global implications. *J. Geophys. Res.* **110**, G01004 (2005).
- Callaghan, T. V. et al. Effects on the function of Arctic ecosystems in the short- and long-term perspectives. *Ambio* **33**, 448–458 (2004).
- Moles, A. T. et al. Global patterns in plant height. *J. Ecol.* **97**, 923–932 (2009).
- Moles, A. T. et al. Global patterns in seed size. *Glob. Ecol. Biogeogr.* **16**, 109–116 (2007).

16. Reich, P. B. & Oleksyn, J. Global patterns of plant leaf N and P in relation to temperature and latitude. *Proc. Natl Acad. Sci. USA* **101**, 11001–11006 (2004).
17. Díaz, S. et al. The global spectrum of plant form and function. *Nature* **529**, 167–171 (2016).
18. Siefert, A. et al. A global meta-analysis of the relative extent of intraspecific trait variation in plant communities. *Ecol. Lett.* **18**, 1406–1419 (2015).
19. McMahon, S. M. et al. Improving assessment and modelling of climate change impacts on global terrestrial biodiversity. *Trends Ecol. Evol.* **26**, 249–259 (2011).
20. Elmendorf, S. C. et al. Experiment, monitoring, and gradient methods used to infer climate change effects on plant communities yield consistent patterns. *Proc. Natl Acad. Sci. USA* **112**, 448–452 (2015).
21. De Frenne, P. et al. Latitudinal gradients as natural laboratories to infer species' responses to temperature. *J. Ecol.* **101**, 784–795 (2013).
22. Sandel, B. et al. Contrasting trait responses in plant communities to experimental and geographic variation in precipitation. *New Phytol.* **188**, 565–575 (2010).
23. Ackerman, D., Griffin, D., Hobbie, S. E. & Finlay, J. C. Arctic shrub growth trajectories differ across soil moisture levels. *Glob. Change Biol.* **23**, 4294–4302 (2017).
24. Wright, I. J. et al. Global climatic drivers of leaf size. *Science* **357**, 917–921 (2017).
25. Wrona, F. J. et al. Transitions in Arctic ecosystems: ecological implications of a changing hydrological regime. *J. Geophys. Res. Biogeosci.* **121**, 650–674 (2016).
26. Read, Q. D., Moorhead, L. C., Swenson, N. G., Bailey, J. K. & Sanders, N. J. Convergent effects of elevation on functional leaf traits within and among species. *Funct. Ecol.* **28**, 37–45 (2014).
27. Albert, C. H., Grassein, F., Schurr, F. M., Vieilledent, G. & Violle, C. When and how should intraspecific variability be considered in trait-based plant ecology? *Perspect. Plant Ecol. Evol. Syst.* **13**, 217–225 (2011).
28. Elmendorf, S. C. et al. Global assessment of experimental climate warming on tundra vegetation: heterogeneity over space and time. *Ecol. Lett.* **15**, 164–175 (2012).
29. Gottfried, M. et al. Continent-wide response of mountain vegetation to climate change. *Nat. Clim. Change* **2**, 111–115 (2012).
30. Blok, D. et al. Shrub expansion may reduce summer permafrost thaw in Siberian tundra. *Glob. Change Biol.* **16**, 1296–1305 (2010).
31. Blok, D., Elberling, B. & Michelsen, A. Initial stages of tundra shrub litter decomposition may be accelerated by deeper winter snow but slowed down by spring warming. *Ecosystems* **19**, 155–169 (2016).
32. Cahoon, S. M. P. et al. Interactions among shrub cover and the soil microclimate may determine future Arctic carbon budgets. *Ecol. Lett.* **15**, 1415–1422 (2012).
33. Lawrence, D. M. & Swenson, S. C. Permafrost response to increasing Arctic shrub abundance depends on the relative influence of shrubs on local soil cooling versus large-scale climate warming. *Environ. Res. Lett.* **6**, 045504 (2011).
34. Christiansen, C. T. et al. Enhanced summer warming reduces fungal decomposer diversity and litter mass loss more strongly in dry than in wet tundra. *Glob. Change Biol.* **23**, 406–420 (2017).
35. Kaarlejärvi, E., Eskelinen, A. & Olofsson, J. Herbivores rescue diversity in warming tundra by modulating trait-dependent species losses and gains. *Nat. Commun.* **8**, 419 (2017).
36. Björkman, A. D., Vellend, M., Frei, E. R. & Henry, G. H. R. Climate adaptation is not enough: warming does not facilitate success of southern tundra plant populations in the high Arctic. *Glob. Change Biol.* **23**, 1540–1551 (2017).
37. Reich, P. B. The world-wide 'fast-slow' plant economics spectrum: a traits manifesto. *J. Ecol.* **102**, 275–301 (2014).
38. Wullschlegel, S. D. et al. Plant functional types in Earth system models: past experiences and future directions for application of dynamic vegetation models in high-latitude ecosystems. *Ann. Bot.* **114**, 1–16 (2014).
39. Butler, E. E. et al. Mapping local and global variability in plant trait distributions. *Proc. Natl Acad. Sci. USA* **114**, E10937–E10946 (2017).
40. Reich, P. B., Rich, R. L., Lu, X., Wang, Y.-P. & Oleksyn, J. Biogeographic variation in evergreen conifer needle longevity and impacts on boreal forest carbon cycle projections. *Proc. Natl Acad. Sci. USA* **111**, 13703–13708 (2014).

Acknowledgements This paper is an outcome of the sTundra working group supported by sDiv, the Synthesis Centre of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig (DFG FZT 118). A.D.B. was supported by an iDiv postdoctoral fellowship and The Danish Council for Independent Research - Natural Sciences (DFG 4181-00565 to S.N.). A.D.B., I.H.M.-S., H.J.D.T. and S.A.-B. were funded by the UK Natural Environment Research Council (ShrubTundra Project NE/M016323/1 to I.H.M.-S.). S.N., A.B.O., S.S.N. and U.A.T. were supported by the Villum Foundation's Young Investigator Programme (VKR023456 to S.N.) and the Carlsberg Foundation (2013-01-0825). N.R. was supported by the DFG-Forschungszentrum 'German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig' and Deutsche Forschungsgemeinschaft DFG (RU 1536/3-1). A.Buc. was supported by EU-F7P INTERACT (262693) and MOBILITY PLUS (1072/MOB/2013/0). A.B.O. was additionally supported by the Danish Council for Independent Research - Natural Sciences (DFG 4181-00565 to S.N.). J.M.A. was supported by the Carl Tryggers stiftelse for vetenskaplig forskning, A.H. by the Research Council of Norway (244557/E50), B.E. and A.Mic. by the Danish National Research Foundation (CENPERM DNRF100), B.M. by the Soil Conservation Service of Iceland and E.R.F. by the Swiss National Science Foundation

(155554). B.C.F. was supported by the Academy of Finland (256991) and JPI Climate (291581). B.J.E. was supported by an NSF ATB, CAREER and Macrosystems award. C.M.I. was supported by the Office of Biological and Environmental Research in the US Department of Energy's Office of Science as part of the Next-Generation Ecosystem Experiments in the Arctic (NGEE Arctic) project. D.B. was supported by The Swedish Research Council (2015-00465) and Marie Skłodowska Curie Actions co-funding (INCA 600398). E.W. was supported by the National Science Foundation (DEB-0415383), UWEC-ORSP and UWEC-BCDT. G.S.-S. and M.I.-G. were supported by the University of Zurich Research Priority Program on Global Change and Biodiversity. H.D.A. was supported by NSF PLR (1623764, 1304040). I.S.J. was supported by the Icelandic Research Fund (70255021) and the University of Iceland Research Fund. J.D.M.S. was supported by the Research Council of Norway (262064). J.S.P. was supported by the US Fish and Wildlife Service. J.C.O. was supported by Klimaat voor ruimte, Dutch national research program Climate Change and Spatial Planning. J.F.J., P.G., G.H.R.H., E.L., N.B.-L., K.A.H., L.S.C. and T.Z. were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). G.H.R.H., N.B.-L., E.L., L.S.C. and L.H. were supported by ArcticNet. G.H.R.H., N.B.-L., M.Tr. and L.S.C. were supported by the Northern Scientific Training Program. G.H.R.H., E.L. and N.B.-L. were additionally supported by the Polar Continental Shelf Program. N.B.-L. was additionally supported by the Fonds de recherche du Québec: Nature et Technologies and the Centre d'études Nordiques. J.P. was supported by the European Research Council Synergy grant SyG-2013-610028 IMBALANCE-P. A.A.-R., O.G. and J.M.N. were supported by the Spanish OAPN (project 534S/2012) and European INTERACT project (262693 Transnational Access). K.D.T. was supported by NSF ANS-1418123. L.E.S. and P.A.W. were supported by the UK Natural Environment Research Council Arctic Terrestrial Ecology Special Topic Programme and Arctic Programme (NE/K000284/1 to P.A.W.). P.A.W. was additionally supported by the European Union Fourth Environment and Climate Framework Programme (Project Number ENV4-CT970586). M.W. was supported by DFG RTG 2010. R.D.H. was supported by the US National Science Foundation. M.J.S. and K.N.S. were supported by the Niwot Ridge LTER (NSF DEB-1637686). H.J.D.T. was funded by a NERC doctoral training partnership grant (NE/L002558/1). V.G.O. was supported by the Russian Science Foundation (14-50-00029). L.B. was supported by NSF ANS (1661723) and S.J.G. by NASA ABoVe (NNX15AU03A/NNX17AE44G). B.B.-L. was supported as part of the Energy Exascale Earth System Model (E3SM) project, funded by the US Department of Energy, Office of Science, Office of Biological and Environmental Research. A.E. was supported by the Academy of Finland (projects 253385 and 297191). E.K. was supported by Swedish Research Council (2015-00498), and S.D. was supported by CONICET, FONCyT and SECYT-UNC, Argentina. The study has been supported by the TRY initiative on plant traits (<http://www.try-db.org>), which is hosted at the Max Planck Institute for Biogeochemistry, Jena, Germany and is currently supported by DIVERSITAS/Future Earth and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. A.D.B. and S.C.E. thank the US National Science Foundation for support to receive training in Bayesian methods (grant 1145200 to N. Thompson Hobbs). We thank H. Bruehlheide and J. Ramirez-Villegas for helpful input at earlier stages of this project. We acknowledge the contributions of S. Mamet, M. Jean, K. Allen, N. Young, J. Lowe, O. Eriksson and many others to trait and community composition data collection, and thank the governments, parks, field stations and local and indigenous people for the opportunity to conduct research on their land.

Reviewer information Nature thanks G. Kunstler, F. Schrodt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.D.B., I.H.M.-S. and S.C.E. conceived the study, with input from the sTundra working group (S.N., N.R., P.S.A.B., A.B.-O., D.B., J.H.C.C., W.C., B.C.F., D.G., S.J.G., K.G., G.H.R.H., R.D.H., J.K., J.S.P., J.H.R.L., C.R., G.S.-S., H.J.D.T., M.V., M.W. and S.Wi.). A.D.B. performed the analyses, with input from I.H.M.-S., N.R., S.C.E. and S.N. D.N.K. made the maps of temperature, moisture and trait change. A.D.B. wrote the manuscript, with input from I.H.M.-S., S.C.E., S.N., N.R. and contributions from all authors. A.D.B. compiled the Tundra Trait Team database, with assistance from I.H.M.-S., H.J.D.T. and S.A.-B. Authorship order was determined as follows: (1) core authors; (2) sTundra participants (alphabetical) and other major contributors; (3) authors contributing both trait (Tundra Trait Team) and community composition (for example, ITEx) data (alphabetical); (4) Tundra Trait Team contributors (alphabetical); (5) contributors who provided community composition data only (alphabetical) and (6) contributors who provided TRY trait data (alphabetical).

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0563-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0563-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anne D. Bjorkman^{1,2,3*}, Isla H. Myers-Smith¹, Sarah C. Elmendorf^{4,5,6}, Signe Normand^{2,7,8}, Nadja Rüger^{9,10}, Pieter S. A. Beck¹¹, Anne Blach-Overgaard^{2,8}, Daan Blok¹², J. Hans C. Cornelissen¹³, Bruce C. Forbes¹⁴, Damien Georges^{1,15}, Scott J. Goetz¹⁶, Kevin C. Guay¹⁷, Gregory H. R. Henry¹⁸, Janneke HilleRisLambers¹⁹, Robert D. Hollister²⁰, Dirk N. Karger²¹, Jens Kattge^{9,22}, Peter Manning³, Janet S. Prevéy²³, Christian Rixen²³, Gabriela Schaepman-Strub²⁴, Haydn J. D. Thomas¹, Mark Vellend²⁵, Martin Wilmsking²⁶, Sonja Wipf²³, Michele Carbognani²⁷, Luise Hermanutz²⁸, Esther Lévesque²⁹, Ulf Molau³⁰, Alessandro Petraglia²⁷, Nadejda A. Soudzilovskaia³¹, Marko J. Spasojevic³², Marcello Tomaselli²⁷, Tage Vowles³³, Juha M. Alatalo³⁴, Heather D. Alexander³⁵, Alba Anadon-Rosell^{26,36,37}, Sandra Angers-Blondin¹, Mariska te Beest^{38,39}, Logan Berner¹⁶, Robert G. Björk^{33,40}, Agata Buchwal^{41,42}, Allan Buras⁴³, Katherine Christie⁴⁴, Elisabeth J. Cooper⁴⁵, Stefan Dullinger⁴⁶, Bo Elberling⁴⁷, Anu Eskelinen^{9,48,49}, Esther R. Frei^{18,21}, Oriol Grau^{50,51}, Paul Grogan⁵², Martin Hallinger⁵³, Karen A. Harper⁵⁴, Monique M. P. D. Heijmans⁵⁵, James Hudson⁵⁶, Karl Hülber⁴⁶, Maitane Iturrate-Garcia²⁴, Colleen M. Iversen⁵⁷, Francesca Jaroszynska^{23,58}, Jill F. Johnstone⁵⁹, Rasmus Halfdan Jørgensen⁶⁰, Elina Kaarlejärvi^{38,61}, Rebecca Klady⁶², Sara Kuleza⁵⁹, Aino Kulonen²³, Laurent J. Lamarque²⁹, Trevor Lantz⁶³, Chelsea J. Little^{24,64}, James D. M. Speed⁶⁵, Anders Michelsen^{47,66}, Ann Milbau⁶⁷, Jacob Nabe-Nielsen⁶⁸, Sigrid Schøler Nielsen², Josep M. Ninot^{36,37}, Steven F. Oberbauer⁶⁹, Johan Olofsson³⁸, Vladimir G. Onipchenko⁷⁰, Sabine B. Rumpf⁴⁶, Philipp Semenchuk^{45,46}, Rohan Shetti²⁶, Laura Siegwart Collier²⁸, Lorna E. Street¹, Katharine N. Suding⁴, Ken D. Tape⁷¹, Andrew Trant^{28,72}, Urs A. Treier^{2,7,8}, Jean-Pierre Tremblay⁷³, Maxime Tremblay²⁹, Susanna Venn⁷⁴, Stef Weijers⁷⁵, Tara Zamin⁵², Noémie Boulanger-Lapointe¹⁸, William A. Gould⁷⁶, David S. Hik⁷⁷, Annika Hofgaard⁷⁸, Ingibjörg S. Jónsdóttir^{79,80}, Janet Jorgenson⁸¹, Julia Klein⁸², Borgthor Magnusson⁸³, Craig Tweedie⁸⁴, Philip A. Wookey⁸⁵, Michael Bahn⁸⁶, Benjamin Blonder^{87,88}, Peter M. van Bodegom⁸⁹, Benjamin Bond-Lamberty⁹⁰, Giandiego Campetella⁹¹, Bruno E. L. Cerabolini⁹², F. Stuart Chapin III⁹³, William K. Cornwell⁹⁴, Joseph Craine⁹⁵, Matteo Dainese⁹⁶, Franciska T. de Vries⁹⁷, Sandra Diaz⁹⁸, Brian J. Enquist^{99,100}, Walton Green¹⁰¹, Ruben Milla¹⁰², Ülo Niinemets¹⁰³, Yusuke Onoda¹⁰⁴, Jenny C. Ordoñez¹⁰⁵, Wim A. Ozinga^{106,107}, Josep Penuelas^{51,108}, Hendrik Poorter^{109,110}, Peter Poschod¹¹¹, Peter B. Reich^{112,113}, Brody Sandel¹¹⁴, Brandon Schamp¹¹⁵, Serge Sheremetev¹¹⁶ & Evan Weiher¹¹⁷

¹School of GeoSciences, University of Edinburgh, Edinburgh, UK. ²Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Aarhus, Denmark. ³Senckenberg Gesellschaft für Naturforschung, Biodiversity and Climate Research Centre (BiK-F), Frankfurt, Germany. ⁴Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. ⁵National Ecological Observatory Network, Boulder, CO, USA. ⁶Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA. ⁷Arctic Research Center, Department of Bioscience, Aarhus University, Aarhus, Denmark. ⁸Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Bioscience, Aarhus University, Aarhus, Denmark. ⁹German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. ¹⁰Smithsonian Tropical Research Institute, Balboa, Panama. ¹¹European Commission, Joint Research Centre, Directorate D — Sustainable Resources, Bio-Economy Unit, Ispra, Italy. ¹²Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden. ¹³Systems Ecology, Department of Ecological Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. ¹⁴Arctic Centre, University of Lapland, Rovaniemi, Finland. ¹⁵International Agency for Research in Cancer, Lyon, France. ¹⁶School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA. ¹⁷Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. ¹⁸Department of Geography, University of British Columbia, Vancouver, British Columbia, Canada. ¹⁹Biology Department, University of Washington, Seattle, WA, USA. ²⁰Biology Department, Grand Valley State University, Allendale, MI, USA. ²¹Swiss Federal Research Institute WSL, Birmensdorf, Switzerland. ²²Max Planck Institute for Biogeochemistry, Jena, Germany. ²³WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland. ²⁴Department of Evolutionary Biology

and Environmental Studies, University of Zurich, Zurich, Switzerland. ²⁵Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada. ²⁶Institute of Botany and Landscape Ecology, Greifswald University, Greifswald, Germany. ²⁷Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy. ²⁸Department of Biology, Memorial University, St. John's, Newfoundland and Labrador, Canada. ²⁹Département des Sciences de l'environnement et Centre d'études nordiques, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada. ³⁰Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden. ³¹Environmental Biology Department, Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands. ³²Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA, USA. ³³Department of Earth Sciences, University of Gothenburg, Gothenburg, Sweden. ³⁴Department of Biological and Environmental Sciences, Qatar University, Doha, Qatar. ³⁵Department of Forestry, Forest and Wildlife Research Center, Mississippi State University, Mississippi State, MS, USA. ³⁶Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona, Barcelona, Spain. ³⁷Biodiversity Research Institute, University of Barcelona, Barcelona, Spain. ³⁸Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden. ³⁹Environmental Sciences, Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands. ⁴⁰Gothenburg Global Biodiversity Centre, Göteborg, Sweden. ⁴¹Institute of Geoeology and Geoinformation, Adam Mickiewicz University, Poznan, Poland. ⁴²Department of Biological Sciences, University of Alaska, Anchorage, Anchorage, AK, USA. ⁴³Forest Ecology and Forest Management, Wageningen University and Research, Wageningen, The Netherlands. ⁴⁴The Alaska Department of Fish and Game, Anchorage, AK, USA. ⁴⁵Department of Arctic and Marine Biology, Faculty of Biosciences, Fisheries and Economics, UiT—The Arctic University of Norway, Tromsø, Norway. ⁴⁶Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria. ⁴⁷Center for Permafrost (CENPERM), Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark. ⁴⁸Department of Physiological Diversity, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany. ⁴⁹Department of Ecology and Genetics, University of Oulu, Oulu, Finland. ⁵⁰Global Ecology Unit, CREA-CSIIC-UAB, Cerdanyola del Vallès, Spain. ⁵¹CREAF, Cerdanyola del Vallès, Spain. ⁵²Department of Biology, Queen's University, Kingston, Ontario, Canada. ⁵³Biology Department, Swedish Agricultural University (SLU), Uppsala, Sweden. ⁵⁴Biology Department, Saint Mary's University, Halifax, Nova Scotia, Canada. ⁵⁵Plant Ecology and Nature Conservation Group, Wageningen University and Research, Wageningen, The Netherlands. ⁵⁶British Columbia Public Service, Surrey, British Columbia, Canada. ⁵⁷Climate Change Science Institute and Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. ⁵⁸Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, UK. ⁵⁹Department of Biology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. ⁶⁰Forest and Landscape College, Department of Geosciences and Natural Resource Management, University of Copenhagen, Nødebo, Denmark. ⁶¹Department of Biology, Vrije Universiteit Brussel (VUB), Brussels, Belgium. ⁶²Department of Forest Resources Management, Faculty of Forestry, University of British Columbia, Vancouver, British Columbia, Canada. ⁶³School of Environmental Studies, University of Victoria, Victoria, British Columbia, Canada. ⁶⁴Department of Aquatic Ecology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland. ⁶⁵NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway. ⁶⁶Department of Biology, University of Copenhagen, Copenhagen, Denmark. ⁶⁷Research Institute for Nature and Forest (INBO), Brussels, Belgium. ⁶⁸Department of Bioscience, Aarhus University, Roskilde, Denmark. ⁶⁹Department of Biological Sciences, Florida International University, Miami, FL, USA. ⁷⁰Department of Geobotany, Lomonosov Moscow State University, Moscow, Russia. ⁷¹Institute of Northern Engineering, University of Alaska Fairbanks, Fairbanks, AK, USA. ⁷²School of Environment, Resources and Sustainability, University of Waterloo, Waterloo, Ontario, Canada. ⁷³Département de biologie, Centre d'études nordiques and Centre d'étude de la forêt, Université Laval, Québec City, Québec, Canada. ⁷⁴Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Burwood, Victoria, Australia. ⁷⁵Department of Geography, University of Bonn, Bonn, Germany. ⁷⁶USDA Forest Service International Institute of Tropical Forestry, Río Piedras, Puerto Rico. ⁷⁷Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada. ⁷⁸Norwegian Institute for Nature Research, Trondheim, Norway. ⁷⁹Faculty of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland. ⁸⁰University Centre in Svalbard, Longyearbyen, Norway. ⁸¹Arctic National Wildlife Refuge, US Fish and Wildlife Service, Fairbanks, AK, USA. ⁸²Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO, USA. ⁸³Icelandic Institute of Natural History, Gardabaer, Iceland. ⁸⁴University of Texas at El Paso, El Paso, TX, USA. ⁸⁵Biology and Environmental Sciences, Faculty of Natural Sciences, University of Stirling, Stirling, UK. ⁸⁶Institute of Ecology, University of Innsbruck, Innsbruck, Austria. ⁸⁷Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK. ⁸⁸Rocky Mountain Biological Laboratory, Crested Butte, CO, USA. ⁸⁹Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands. ⁹⁰Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA. ⁹¹School of Biosciences and Veterinary Medicine, Plant Diversity and Ecosystems Management Unit, University of Camerino, Camerino, Italy. ⁹²DiSTA, University of Insubria, Varese, Italy. ⁹³Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK, USA. ⁹⁴School of Biological, Earth and Environmental Sciences, Ecology and Evolution Research Centre, UNSW Sydney, Sydney, New South Wales, Australia. ⁹⁵Jonah Ventures, Boulder, CO, USA. ⁹⁶Institute for Alpine Environment, Eurac Research, Bolzano, Italy. ⁹⁷School of Earth and Environmental Sciences, The University of Manchester, Manchester, UK. ⁹⁸Instituto Multidisciplinario de Biología Vegetal (IMBIV), CONICET and FCEyN, Universidad Nacional de Córdoba, Córdoba, Argentina. ⁹⁹Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. ¹⁰⁰The Santa Fe Institute, Santa Fe, NM, USA. ¹⁰¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. ¹⁰²Área de Biodiversidad y Conservación. Departamento de Biología, Geología, Física y Química Inorgánica, Universidad Rey Juan Carlos, Madrid, Spain. ¹⁰³Estonian University of Life Sciences, Tartu, Estonia. ¹⁰⁴Graduate School of Agriculture, Kyoto University,

Kyoto, Japan. ¹⁰⁵World Agroforestry Centre — Latin America, Lima, Peru. ¹⁰⁶Team Vegetation, Forest and Landscape Ecology, Wageningen Environmental Research (Alterra), Wageningen, The Netherlands. ¹⁰⁷Institute for Water and Wetland Research, Radboud University Nijmegen, Nijmegen, The Netherlands. ¹⁰⁸Global Ecology Unit CREA-CSIU-UAB, Consejo Superior de Investigaciones Científicas, Bellaterra, Spain. ¹⁰⁹Plant Sciences (IBG-2), Forschungszentrum Jülich GmbH, Jülich, Germany. ¹¹⁰Department of Biological Sciences, Macquarie University, North Ryde, New South Wales, Australia. ¹¹¹Ecology and Conservation Biology, Institute of Plant

Sciences, University of Regensburg, Regensburg, Germany. ¹¹²Department of Forest Resources, University of Minnesota, St. Paul, MN, USA. ¹¹³Hawkesbury Institute for the Environment, Western Sydney University, Penrith, New South Wales, Australia. ¹¹⁴Department of Biology, Santa Clara University, Santa Clara, CA, USA. ¹¹⁵Department of Biology, Algoma University, Sault Ste. Marie, Ontario, Canada. ¹¹⁶Komarov Botanical Institute, St Petersburg, Russia. ¹¹⁷Department of Biology, University of Wisconsin — Eau Claire, Eau Claire, WI, USA. *e-mail: anne.bjorkman@senckenberg.de

METHODS

Below we describe the data, workflow (Extended Data Fig. 1b) and detailed methods used to conduct all analyses. No statistical methods were used to predetermine sample size.

Community composition data. Community composition data used for calculating CWM were compiled from a previous synthesis of tundra vegetation resurveys² (including many International Tundra Experiment (ITEX) sites) and expanded with additional sites (for example, Gavia Pass in the Italian Alps and three sites in Sweden) and years (for example, 2015 survey data added for Iceland sites, Qikiqtaruk–Herschel Island (QHI) and Alexandra Fiord; Supplementary Table 2). We included only sites for which community composition data were roughly equivalent to percentage cover (that is, excluding estimates approximating biomass), for a total of 117 sites (defined as plots in a single contiguous vegetation type) within 38 regions (defined as a CRU⁴¹ grid cell). Plot-level surveys of species composition and cover were conducted at each of these sites between 1989 and 2015 (see the previous study² for more details regarding data collection and processing). On average, there were 15.2 plots per site. Repeat surveys were conducted over a minimum duration of 5 and up to 21 years between 1989 and 2015 (mean duration, 13.6 years), for a total of 1,781 unique plots and 5,507 plot–year combinations. Plots were either permanent (that is, staked; 62% of sites) or semi-permanent (38%), such that the approximate but not exact location was resurveyed. The vegetation monitoring sites were located in treeless Arctic or alpine tundra and ranged in latitude from 40° (Colorado Rockies) to 80° (Ellesmere Island, Canada) and were circumpolar in distribution (Fig. 1a and Supplementary Table 2). Our analyses only include vascular plants, because there was insufficient trait data for non-vascular species. Changes in bryophytes and other cryptogams are an important part of the trait and function change in tundra ecosystems^{42,43}, thus the incorporation of non-vascular plants and their traits is a future research priority.

Temperature extraction for community composition observations. We extracted summer (warmest quarter) and winter (coldest quarter) temperature estimates for each of the vegetation survey sites from both the WorldClim⁴⁴ (for long-term averages; <http://www.worldclim.org/>) and CRU⁴¹ (for temporal trends; <http://www.cru.uea.ac.uk/>) gridded climate datasets. WorldClim temperatures were further corrected for elevation (based on the difference between the recorded elevation of a site and the mean elevation of the WorldClim grid cell) according to a correction factor of -0.005°C per m increase in elevation. This correction factor was calculated by extracting the mean temperature and elevation (WorldClim 30-s resolution maps) of all cells that fall in a 2.5-km radius buffer around our sites and fitting a linear mixed model (with site as a random effect) to estimate the rate of temperature change with elevation.

The average long-term (1960–present) temperature trend across all sites was 0.26°C (range, -0.06 to 0.49) and 0.43°C (range, -0.15 to 1.32) per decade for summer and winter temperature, respectively.

Soil moisture for community composition observations. A categorical measure of soil moisture at each site was provided by the principle investigator of the site according to previously described methods^{2,45}. Soil moisture was considered to be (1) dry when during the warmest month of the year the top 2 cm of the soil was dry to the touch; (2) mesic when soils were moist year round, but standing water was not present; and (3) wet when standing water was present during the warmest month of the year.

Soil moisture change for maps of environmental and trait change. We used high-resolution observations of soil moisture from the European Space Agency (ESA) CCI SM v.04.2 to estimate soil moisture change over time (Fig. 4c). To calculate the mean distribution of soil moisture, we averaged the observations for the period between 1979 and 2016. Because the ESA CCI SM temporal coverage is poor for our sites, temporal data were instead taken from the European Re-analysis (ERA-Interim; volumetric soil water layer 1) soil moisture estimates for the same time period. We downsampled the ERA-Interim data to the 0.05° resolution of ESA CCI SM v.04.2 using climatologically aided interpolation (delta change method)⁴⁶. The change in soil water content was then calculated separately for each grid cell using linear regression with month as a predictor variable. To classify the soil moisture data into three categories (wet, mesic or dry) to match the community composition dataset, we used a quantile approach on the mean soil moisture within the extent of the Arctic. We assigned the lowest quantile to dry and the highest to wet conditions. For the trends in soil moisture between 1979 and 2016, we first calculated the percentage change in relation to the mean, and then calculated the change based on the categorical data (for example, 5% change from category 1 (dry) to category 2 (mesic)).

Changes in water availability for analysis. Although the strong effect of soil moisture on spatial temperature–trait relationships suggests that change in water availability over time will play an important part in mediating trait change, we did not use the CRU estimates of precipitation change over time, because of issues with precipitation records at high latitudes and the inability of gridded datasets to capture localized precipitation patterns^{47,48}. The CRU precipitation trends at our sites

included many data gaps filled by long-term mean values, especially at high-latitude sites⁴⁵. As a purely exploratory analysis, we used the downsampled ERA-Interim data described above to investigate whether trait change is related to summer soil moisture change (June, July and August; Extended Data Fig. 5b). However, we caution that changes in soil moisture in our tundra sites are primarily controlled by the timing of the snow melting, soil drainage, the permafrost table and local hydrology²⁵, and as such precipitation records and coarse-grain remotely sensed soil-moisture change data are unlikely to accurately represent local changes in soil water availability. For this reason, we did not use the ERA-Interim data to explore spatial relationships between temperature, moisture and community traits, as the categorical soil moisture data (described above) were collected specifically within each community composition site and are therefore a more accurate representation of long-term mean soil moisture conditions in that specific location.

Trait data. Continuous trait data (adult plant height, leaf area (average one-sided area of a single leaf), SLA (leaf area per unit of leaf dry mass), leaf nitrogen content (per unit of leaf dry mass), and LDMC (leaf dry mass per unit of leaf fresh mass) (Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1) were extracted from the TRY⁴⁹ 3.0 database (<https://www.try-db.org/TryWeb/Home.php>). We also ran a field and data campaign in 2014–2015 to collect additional in situ tundra trait data (the ‘Tundra Trait Team’ (TTT) dataset⁵⁰) to supplement existing TRY records. All species names from the vegetation monitoring sites, TRY and TTT were matched to accepted names in The Plant List using the R package Taxonstand⁵¹ (v.1.8) before merging the datasets. Community-level traits (woodiness and evergreenness) were derived from functional group classifications for each species². Woodiness is estimated as the proportion (abundance) of woody species in the plot, whereas evergreenness is the proportion of evergreen woody species abundance out of all woody species (evergreen plus deciduous) in a plot. Because some sites did not contain any woody species (and thus the proportion of evergreen woody species could not be calculated), this trait is only estimated for 98 of the 117 total sites.

Data cleaning for TRY data. TRY trait data were subjected to a multi-step cleaning process. First, all values that did not represent individual measurements or approximate species means were excluded. When a dataset within TRY contained only coarse plant height estimates (for example, estimated to the nearest foot), we removed these values unless no other estimate of height for that species was available. We then identified overlapping datasets within TRY and removed duplicate observations whenever possible. The following datasets were identified as having partially overlapping observations: GLOPNET (Global Plant Trait Network Database), The LEDA Traitbase, Abisko and Sheffield Database, Tundra Plant Traits Database and Kew Seed Information Database (SID).

We then removed duplicates within each TRY dataset (for example, if a value is listed once as ‘mean’ and again as ‘best estimate’) by first calculating the ratio of duplicated values within each dataset, and then removing duplicates from datasets with more than 30% duplicated values. This cut-off was determined by manual evaluation of datasets at a range of thresholds. Datasets with fewer than 30% duplicated values were not removed in this way as any internally duplicate values were assumed to be true duplicates (that is, two different individuals were measured and happened to have the same measurement value).

We also removed all species mean observations from the ‘Niwot Alpine Plant Traits’ database and replaced it with the original individual observations provided by M.J.S.

Data cleaning for the combined TRY and TTT dataset. Both datasets were checked for improbable values, with the goal of excluding likely errors or measurements with incorrect units but without excluding true extreme values. We followed a series of data-cleaning steps, in each case identifying whether a given observation (x) was likely to be erroneous (that is, ‘error risk’) by calculating the difference between x and the mean (excluding x) of the taxon and then dividing by the standard deviation of the taxon.

We used a hierarchical data-cleaning method, because the standard deviation of a trait value is related to the mean and sample size. First, we checked individual records against the entire distribution of observations of that trait and removed any records with an error risk greater than 8 (that is, a value more than 8 standard deviations away from the trait mean). For species that occurred in four or more unique datasets within TRY or TTT (that is, different data contributors), we estimated a species mean per dataset and removed observations for which the species mean error risk was greater than 3 (that is, the species mean of that dataset was more than 3 standard deviations away from the species mean across all datasets). For species that occurred in fewer than four unique datasets, we estimated a genus mean per dataset and removed observations in datasets for which the error risk based on the genus mean was greater than 3.5. Finally, we compared individual records directly to the distribution of values for that species. For species with more than four records, we excluded values above an error risk Y , where Y was dependent on the number of records of that species and ranged from an error risk of 2.25 for species with fewer than 10 records to an error risk of 4 for species with more than 30 records. For species with four or fewer records, we manually checked trait

values and excluded only those that were obviously erroneous, based on our expert knowledge of these species.

This procedure was performed on the complete tundra trait database, including species and traits not presented here. In total 2,056 observations (1.6%) were removed. In all cases, we visually checked the excluded values against the distribution of all observations for each species to ensure that our trait cleaning protocol was reasonable.

Trait data were distributed across latitudes within the tundra biome (Extended Data Fig. 1a). All trait observations with latitude and longitude information were mapped and checked for implausible values (for example, falling in the ocean). These values were corrected from the original publications or by contacting the data contributor whenever possible.

Final trait database. After removing duplicates and outliers as described above, we retained 56,048 unique trait observations (of which 18,613 are contained in TRY and 37,435 were newly contributed by the TTT⁵⁰ field campaign) across the five continuous traits of interest. Of the 447 identified species in the ITEX dataset, 386 (86%) had trait data available from TRY or TTT for at least one trait (range 52–100% per site). Those species without trait data generally represent rare or uncommon species unique to each site; on average, trait data were available for 97% of total plant cover across all sites (range 39–100% per site; Supplementary Table 1).

Temperature extraction for trait observations. WorldClim climate variables were extracted for all trait observations with latitude and longitude values recorded (53,123 records in total, of which 12,380 were from TRY and 33,621 from TTT). Because most observations did not include information about elevation, temperature estimates for individual trait observations were not corrected for elevation and thus represent the temperature at the mean elevation of the WorldClim grid cell.

Analyses. Terminology. Here we provide a brief description of acronyms and symbols used in the methods and model equations. α is used to designate lower-level model intercepts; β is used to designate lower-level model slopes; γ is used to designate the model parameters of interest (for example, the temperature–trait relationship); CWM designates the mean trait value of all species in a plot, weighted by their abundance in the plot; CWM + ITV designates CWM adjusted with the estimated contribution of ITV based on the intraspecific temperature–trait relationship of each species; and ITV designates variation in trait values within the same species (that is, intraspecific trait variation).

Models. All analyses were conducted in JAGS and/or Stan through R (v.3.3.3) using packages rjags⁵² (v.4.6) and rstan⁵³ (v.2.14.1). In all cases, models were run until convergence was reached, which was assessed both visually in traceplots and by ensuring that all Gelman–Rubin convergence diagnostic (\hat{R})⁵⁴ values were less than 1.1.

A major limitation of the species mean trait approach, which is often used in analyses of environment–trait relationships, has been the failure to account for ITV, which could be as or more important than interspecific variation^{55,56}. We addressed this issue by using a hierarchical analysis that incorporates both within-species and community-level trait variation across climate gradients to estimate trait change over space and time at the biome scale. We used a Bayesian approach that accounts for the hierarchical spatial (plots within sites within regions) and taxonomic (intra- and inter-specific variation) structure of the data as well as uncertainty in estimated parameters introduced through absences in trait records for some species, or through taxa that were identified to genus or functional group (rather than species) in vegetation surveys.

ITV. To calculate intraspecific temperature–trait relationships, we used a subset of the trait dataset containing only those species for which traits had been measured in at least four unique locations spanning a temperature range of at least 10% of the entire temperature range (2.6°C and 5.0°C for summer and winter temperature, respectively), and for which the latitude and longitude of the measured individual or group of individuals was recorded. The number of species meeting these criteria varied by trait and temperature variable: 108 and 109 for SLA, 80 and 86 for plant height, 74 and 72 for leaf nitrogen, 85 and 76 for leaf area, and 43 and 52 for LDMC, for summer and winter temperature, respectively. These species counts correspond to 53–73% of the abundance in the community. The relationship between each trait and temperature (Fig. 2b) was estimated from a Bayesian hierarchical model, with temperature as the predictor variable and species (sp) and dataset-by-location (d) modelled as random effects:

$$\text{trait}_i^{\text{observed}} \sim \log \text{normal}(\alpha_{\text{sp},d}, \sigma_{\text{sp}})$$

$$\alpha_{\text{sp},d} \sim \text{Normal}(\alpha_{\text{sp}} + \beta_{\text{sp}} T_d, \sigma_1)$$

$$\beta_{\text{sp}} \sim \text{Normal}(B, \sigma_2)$$

$$\alpha_{\text{sp}} \sim \text{Normal}(A, \sigma_3)$$

in which the tilde (\sim) indicates ‘distributed as’, T indicates temperature, i represents each trait observation and A and B are the intercept and slope hyperparameters, respectively. Because LDMC represents a ratio and is thus bound between 0 and 1, we used a beta error distribution for this trait. Temperature values were mean-centred within each species. We used non-informative priors for all coefficients.

We further explored whether the strength of intraspecific temperature–height relationships varied by functional group. We find that all functional groups (including dwarf shrubs, which are genetically limited in their ability to grow upright) show similar temperature–trait relationships (Extended Data Fig. 9a). These results suggest that intraspecific temperature–height relationships are not just a consequence of individual growth differences, and are not restricted to particular functional groups with greater capacity for vertical growth (for example, tall shrubs and graminoids versus dwarf shrubs and certain forb species).

Calculation of CWM values. We calculated the CWM (that is, the mean trait value of all species in a plot, weighted by the abundance of each species), for all plots within a site. We used a Bayesian approach to calculate trait means for every species (s) using an intercept-only model (such that the intercept per species (α_s) is equivalent to the mean trait value of the species) and variation per species (σ_s) with a lognormal error distribution:

$$\text{trait}_i^{\text{observed}} \sim \log \text{normal}(\alpha_s, \sigma_s)$$

Because LDMC represents a ratio and is thus bound between 0 and 1, we used a beta error distribution instead of log normal for this trait. When a species was measured multiple times in several different locations, we additionally included a random effect of dataset-by-location (d) to reduce the influence of a single dataset with many observations at one site when calculating the mean per species:

$$\text{trait}_i^{\text{observed}} \sim \log \text{normal}(\alpha_{s,d}, \sigma_d)$$

$$\alpha_{s,d} \sim \text{Normal}(\alpha_s, \sigma_s)$$

We used non-informative priors for all species intercept parameters for which there were four or more unique trait observations, so that the species-level intercept and variance around the intercept per species were estimated from the data. To avoid removing species with little or no trait data from the analyses, we additionally used a ‘gap-filling’ approach that enabled us to estimate the trait mean of each species while accounting for uncertainty in the estimation of this mean. For species with fewer than four but more than one trait observation, we used a normal prior with the mean equal to the mean of the observation(s) and variance estimated based on the mean mean:variance ratio across all species. In other words, we calculated the ratio of mean trait values to the standard deviation of those trait values per species for all species with greater than four observations, then took the mean of these ratios across all species and multiplied this number by the mean of species X (in which X is a species with 1–4 observations) to get the prior for σ . For species with no observations (see Supplementary Table 1), we used a prior mean equal to the mean of all species in the same genus and a prior variance estimated based on the mean mean:variance ratio of all species in that genus or $1.5 \times$ the mean, whichever was lower. If there were no other species in the same genus, then we used a prior mean equal to the mean of all other species in the family and a prior variance estimated based on the mean mean:variance ratio of all species in the family or $1.5 \times$ the mean, whichever was lower.

Incorporating uncertainty in species traits to calculate CWM values. To include uncertainty about species trait means (owing to ITV, missing trait information for some species or when taxa were identified to genus or functional group rather than species) in subsequent analyses, we estimated community-level trait values per plot by sampling from the posterior distribution (mean \pm s.d.) of each species intercept estimate and multiplying this distribution by the relative abundance of each species in the plot to get a CWM distribution per plot (p) per year (y):

$$\text{Normal}\left(\text{CWM}_{p,y}^{\text{mean}}, \text{CWM}_{p,y}^{\text{s.d.}}\right)$$

This approach generates a distribution of CWM values per plot that propagates the uncertainty in the mean estimate of each trait for each species into the plot-level (CWM) estimate. By using a Bayesian approach, we are able to carry through uncertainty in mean estimates of traits to all subsequent analyses and reduce the potential for biased or deceptively precise estimates due to missing trait observations.

Partitioning turnover and estimating contribution of ITV to temperature–trait relationships. To assess the degree to which the spatial temperature–trait relationships are caused by species turnover versus shifts in abundance among sites, we repeated each analysis using the non-weighted community mean (all species weighted equally) of each plot. Temperature–trait relationships estimated with non-weighted community means are due only to species turnover across sites.

Finally, we assessed the potential contribution of ITV to the community-level temperature–trait relationship by using the modelled intraspecific temperature–trait relationship (see ‘ITV’) to predict trait ‘anomaly’ values for each species at each site based on the temperature of that site in a given year relative to its long-term average. An intraspecific temperature–trait relationship could not be estimated for every species owing to an insufficient number of observations for some species. Therefore, we used the mean intraspecific temperature–trait slope across all species to predict trait anomalies for species without intraspecific temperature–trait relationships.

Site- and year-specific species trait estimates were then used to calculate ITV-adjusted CWM (CWM + ITV) for each plot in each measured year, and modelled as for CWM alone. As these adjusted values are estimated relative to the mean value of each species, the spatial temperature–trait relationship that includes this adjustment does not remove any bias in the underlying species mean data. For example, if southern tundra species tend to be measured at the southern edge of their range while northern tundra species tend to be measured at the northern edge of their range, the overall spatial temperature–trait relationship could appear stronger than it really is for species with temperature-related intraspecific variation. This is a limitation of any species-mean approach.

Estimates of temporal CWM + ITV temperature–trait relationships are not prone to this same limitation as they represent relative change, but should also be interpreted with caution as intraspecific temperature–trait relationships may be due to genetic differences among populations rather than plasticity, thus suggesting that trait change would not occur immediately with warming. We therefore caution that the CWM + ITV analyses represent estimates of the potential contribution of ITV to overall CWM temperature–trait relationships over space and time, but should not be interpreted as measured responses.

In summary, we incorporate intraspecific variation into our analyses in three ways. First, by using the posterior distribution (rather than a single mean value) of species trait mean estimates in our calculations of CWM values per plot, so that information about the amount of variation within species is incorporated into all the analyses in our study. Second, by explicitly estimating intraspecific temperature–trait relationships based on the spatial variation in traits among individuals of the same species. Third, by using these modelled temperature–trait relationships to inform estimates of the potential contribution of ITV to overall (CWM + ITV) temperature–trait relationships over space and time.

Spatial community trait models. To investigate spatial relationships in plant traits with summer or winter temperature and soil moisture (Fig. 2a, c), we used a Bayesian hierarchical modelling approach in which soil moisture and soil moisture \times temperature vary at the site level while temperature varies by WorldClim region (unique WorldClim grid \times elevation groups). In total, there were 117 sites (s) nested within 73 WorldClim regions (r). We used only the first year of survey data at each site to estimate spatial relationships in community traits.

$$\text{CWM}_p^{\text{mean}} \sim \text{Normal}(\alpha_s + \alpha_r, \text{CWM}_p^{\text{s.d.}})$$

$$\alpha_s \sim \text{Normal}(\gamma_1 M_s + \gamma_2 M_s T_s, \sigma_1)$$

$$\alpha_r \sim \text{Normal}(\gamma_0 + \gamma_3 T_r, \sigma_2)$$

in which M indicates moisture, $\text{CWM}_p^{\text{mean}}$ is the mean of the posterior distribution of the CWM estimate per plot (p) and $\text{CWM}_p^{\text{s.d.}}$ is the standard deviation of the posterior distribution of the CWM estimate per plot (see ‘Incorporating uncertainty in species traits to calculate CWM values’). See Supplementary Information for complete STAN code.

As woodiness and evergreenness represent proportional data (bound between 0 and 1, inclusive), we used a beta–Bernoulli mixture model of the same structure as above to estimate trait–temperature–moisture relationships for these traits (Extended Data Fig. 3a, b). The discrete and continuous components of the data were modelled separately, with mixing occurring at the site- and region-level estimates (α_s and α_r).

Because Arctic and alpine tundra sites might differ in their trait–environment relationships owing to environmental differences (for example, in soil drainage), we also performed a version of the spatial community trait analyses in which the elevation of each site is visually indicated (not modelled; Extended Data Fig. 9b). We did not attempt to separately analyse trait–environment relationships for Arctic and alpine sites owing to the ambiguity in defining this cut-off (that is, many sites can be categorized as both Arctic and alpine, particularly in Scandinavia and Iceland) and because of the small number of southern, high-alpine sites (European Alps and Colorado Rockies).

For estimation of the overall temperature–trait relationship, we used a model structure similar to that above but with only temperature as a predictor (that is,

without soil moisture). This model was used for both CWM and non-weighted mean estimates to determine the degree to which temperature–trait relationships over space are due to species turnover alone (non-weighted mean) and for CWM + ITV plot-level estimates to determine the likely additional contribution of ITV to the overall temperature–trait relationship, as described above.

Standardized effect sizes for CWM temperature–trait relationships (Fig. 2c) were obtained by dividing the slope of the temperature–trait relationship by the standard deviation of the CWM model residuals. Effect sizes for ITV, turnover only and CWM + ITV were estimated relative to the CWM value for that same trait based on the slope values of each temperature–trait relationship.

Trait change over time. Change over time (Fig. 3a, b) was modelled at the CRU grid cell (region) level. We first estimated a region-by-year (r, y) effect to account for the non-independence of observations made within the same region and year. We also included site (s) as a random effect when there was more than one site per region (to account for non-independence of sites within a region) and plot (p) as a random effect for those sites with permanent (repeating) plots (to account for repeated measures on the same plot over time):

$$\text{CWM}_{p,y}^{\text{mean}} \sim \text{Normal}(\alpha_p + \alpha_s + \alpha_{r,y}, \text{CWM}_{p,y}^{\text{s.d.}})$$

in which $\text{CWM}_{p,y}^{\text{mean}}$ is the mean of the posterior distribution of the CWM estimate per plot (p) in a given year (y) and $\text{CWM}_{p,y}^{\text{s.d.}}$ is the standard deviation of the posterior distribution of the CWM estimate per plot and year (see ‘Incorporating uncertainty in species traits to calculate CWM values’). For non-permanent plots and for sites that were the only site within a region, α_p or α_s , respectively, were set to 0.

Region-level slopes were then used to fit an average trend of community trait values over time, in which Y represents calendar year (centred within each region) as a linear predictor:

$$\alpha_{r,y} \sim \text{Normal}(\alpha_r + \beta_r Y_{r,y}, \sigma_0)$$

$$\beta_r \sim \text{Normal}(B, \sigma_1)$$

$$\alpha_r \sim \text{Normal}(A, \sigma_2)$$

in which A and B are the intercept and slope hyperparameters, respectively. See Supplementary Information for complete STAN code. This model was used for both CWM and non-weighted mean plot-level estimates to determine the degree to which temporal trait change is due to species turnover alone (non-weighted mean) and for CWM + ITV plot-level estimates to determine the potential additional contribution of ITV to overall trait change. We did not account for temporal autocorrelation in these models as most plots were not measured annually (average survey interval = 7.2 years) and did not have more than three observations over the study period (average number of survey years per plot = 3.1).

Standardized effect sizes for CWM change over time (Fig. 3b) were obtained by dividing the slope of overall trait change over time (mean hyperparameter across 117 sites) by the standard deviation of the slope estimates per site. Effect sizes for turnover-only and CWM + ITV changes are estimated relative to the CWM change value for that trait based on the slope values of each.

To estimate the change in the proportion of woody and evergreen species over time (CWM change only; Extended Data Fig. 3c, d) we used a beta–Bernoulli mixture model of the same form described above. The discrete and continuous components of the data were modelled separately, with mixing occurring at the region \times year effect ($\alpha_{r,y}$). We additionally assessed whether the rate of observed trait change over time was related to the duration of vegetation monitoring at each site. There was no influence of monitoring duration for any trait (data not shown).

Temperature sensitivity. Temperature sensitivity (Fig. 3c) was modelled as the variation in CWM trait values with variation in the five-year mean temperature (that is, the mean temperature of the survey year and the four preceding years). A four-year lag was chosen because this interval has been shown to best explain vegetation change in tundra²⁰ and alpine²⁹ plant communities. The model specifics are exactly as shown above (see ‘Trait change over time’), but with temperature in the place of the linear year predictor (Y). Temperatures were centred within each region.

Observed versus expected changes in traits. To compare rates of observed versus expected community trait change (Fig. 4a), we first calculated the mean rate of temperature change across the 38 regions in our study, and then estimated the expected degree of change in each trait over the same period based on this temperature change and the spatial relationship between temperature and CWM trait values (see ‘Spatial community trait models’). We then compared this expected trait change to actual trait change over time (see ‘Trait change over time’). To create Fig. 4a, we used the overall predicted mean value of each trait in the first year of survey (1989) as an intercept, and then used the expected and observed rates of trait change (\pm uncertainty) to predict community trait values in each year

thereafter. We subtracted the intercept from all predicted values to show trait change as an anomaly (difference from 0). The difference between the expected (black) and observed (coloured) lines in Fig. 4a represents a deviation from expected. To calculate total trait change, including the estimated contribution of intraspecific change (coloured dashed lines), we followed the same procedure as described for 'observed' trait change but where this observed change was based on plot-level CWM + ITV estimates that varied by year based on the temperature in that year and the temperature–trait relationship per species (see 'Partitioning turnover and estimating contribution of ITV to temperature–trait relationships'). *Trait change versus changes in temperature and soil moisture.* To determine whether the rate of trait change can be explained by the rate of temperature change at a site, the (static) level of soil moisture of a site or their interaction (Extended Data Fig. 5), we modelled the rate of trait change (see 'Trait change over time') and compared it to the rate of temperature change over the same time interval (with a lag of four years) and soil moisture:

$$\beta_r \sim \text{Normal}(\gamma_0 + \gamma_1 T_r + \gamma_2 M_r + \gamma_3 T_r M_r, \sigma)$$

in which β_r is the rate of trait change per region (Extended Data Fig. 5a). When sites within a region were measured over different intervals or contained different soil moisture estimates, they were modelled separately to match with temperature change estimates over the same interval and soil moisture estimates, which varied at the site level.

We also conducted this analysis using estimates of soil moisture change (with a lag of four years) from downscaled ERA-Interim data (volumetric soil water layer 1). This model took the same form as above, but with moisture change in place of static soil moisture estimates (Extended Data Fig. 5b). Trait change was modelled at the site (rather than region) level, because estimates of soil moisture change varied at the site level. Because ERA-Interim data were not available for every site, this analysis was conducted with a total of 101 rather than 117 sites. We note that the results of this analysis should be interpreted with caution, as local changes in soil moisture may not be well-represented by coarse-scale remotely sensed data, as previously described.

Species gains and losses as a function of traits. We explored whether turnover in community composition was related to species' traits (Extended Data Fig. 6). We estimated species gains and losses at the site (rather than plot) level to reduce the effect of random fluctuations in species presences and absences due to observer error or sampling methodology. Thus, sites with repeating and non-repeating plots were treated the same. A 'gain' was defined as a species that did not occur in a site in the first survey year but did in the last survey year, whereas a 'loss' was the reverse. We then modelled the probability of gain or loss separately as a function of the mean trait value of each species. For example, for gains, all newly observed species received a response type of 1 whereas all other species in the site received a response type of 0:

$$\text{response}_i \sim \text{Bernoulli}(\alpha_s + \alpha_r + \beta_i \text{trait}_i)$$

$$\alpha_r \sim \text{Normal}(A, \sigma_1)$$

$$\beta_i \sim \text{Normal}(B, \sigma_2)$$

$$\alpha_s \sim \text{Normal}(0, \sigma_r)$$

We included a random effect for site (s) only when there were multiple sites within the same region (r), otherwise α_s was set to 0. We considered the responses of species to be related to a given trait when the 95% credible interval on the slope hyperparameter (B) did not overlap zero.

Trait projections with warming. We projected trait change for the minimum (Representative Concentration Pathways (RCP)2.6) and maximum (RCP8.5) IPCC carbon emission scenarios from the NIMR HadGEM2-AO Global Circulation Model (Extended Data Fig. 7). We used the midpoint years of the WorldClim (1975) and HadGem2 (2090) estimates to calculate the expected rate of temperature change over this time period. We then predicted trait values for each year into the future based on the projected rate of temperature change and the spatial relationship between temperature and community trait values (see 'Spatial community trait models').

These projections are not intended to predict actual expected changes in traits over the next century, as many other factors not accounted for here will also influence this change. In particular, future changes in functional traits will probably depend on concurrent changes in moisture availability, which are less well understood than temperature change. Recent modelling efforts predict increases in precipitation across much of the Arctic⁵⁷, but it is unknown whether increasing precipitation will also lead to an increase in soil moisture and/or water availability for plants, as the drying effect of warmer temperatures (for example, due to

increased evaporation and/or decreased duration of snow cover⁵⁸) may outweigh the effect of increased precipitation. Instead, these projections are an attempt to explore theoretical changes in traits over a long-term period when using a space-for-time substitution approach.

Principal component analysis. We performed an ordination of CWM values per plot on all seven traits (Extended Data Fig. 8). Because community evergreenness could only be estimated for plots with at least one woody species, the total number of plots included in this analysis is reduced compared to the entire dataset (1,098 plots out of 1,520 in total). We used the R package *vegan*⁵⁹ (v.2.4.6) to conduct a principal component analysis (PCA) of these data. This analysis uses only trait means per plot, and therefore information about CWM uncertainty due to ITV and/or missing species is lost. The analysis was performed on log-transformed trait values⁴⁹. We extracted the axis coordinates of each plot from the PCA and used the spatial trait–temperature–moisture model described above (see 'Spatial community trait models') to determine whether plot positions along both PCA axes varied with temperature, moisture and their interaction.

Trends in the abundance of species. To provide more insights into the species-specific changes that have occurred over time in tundra ecosystems, we calculated trends in abundance for the most common (widespread and abundant) species in the community composition dataset (Supplementary Table 10). We estimated trends for all species that occurred in at least five sites at a minimum abundance of 5% cover (mean of all plots within a site) across all years. We additionally included species that occurred at low abundance (1% or more) but were widespread (at least 10 sites). This technique yielded a total of 79 species. Abundance changes were modelled as described for trait change over time, but because abundance (proportion of plot cover) is bounded between 0 and 1, inclusive, we used a beta–Bernoulli mixture model. Abundance change was then estimated per species (sp) across all regions (r):

$$\alpha_{sp,r,y} \sim \text{Normal}(\alpha_{sp,r} + \beta_{sp,r} Y_{sp,r,y}, \sigma_{sp})$$

$$\beta_{sp,r} \sim \text{Normal}(B_{sp}, \sigma_1)$$

$$\alpha_{sp,r} \sim \text{Normal}(A_{sp}, \sigma_2)$$

We additionally extracted region-specific slopes per species ($\beta_{sp,r}$) to calculate a proportion of regions in which a given species was increasing or decreasing ('Prop. Increase' and 'Prop. Decrease' in Supplementary Table 10). Because regional slopes are modelled as random effects, these estimates are not entirely independent (that is, they will be pulled towards the overall species mean slope), but provide an approximate estimate of whether directional trends in abundance are consistent across the range of a species.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. STAN code for the two main models (spatial temperature–moisture–trait relationships and community trait change over time) is provided in the Supplementary Information. Code for trait data cleaning is provided in the Tundra Trait Team data repository⁵⁰ (<https://github.com/ShrubHub/TraitHub>, <https://tundratraitteam.github.io/>).

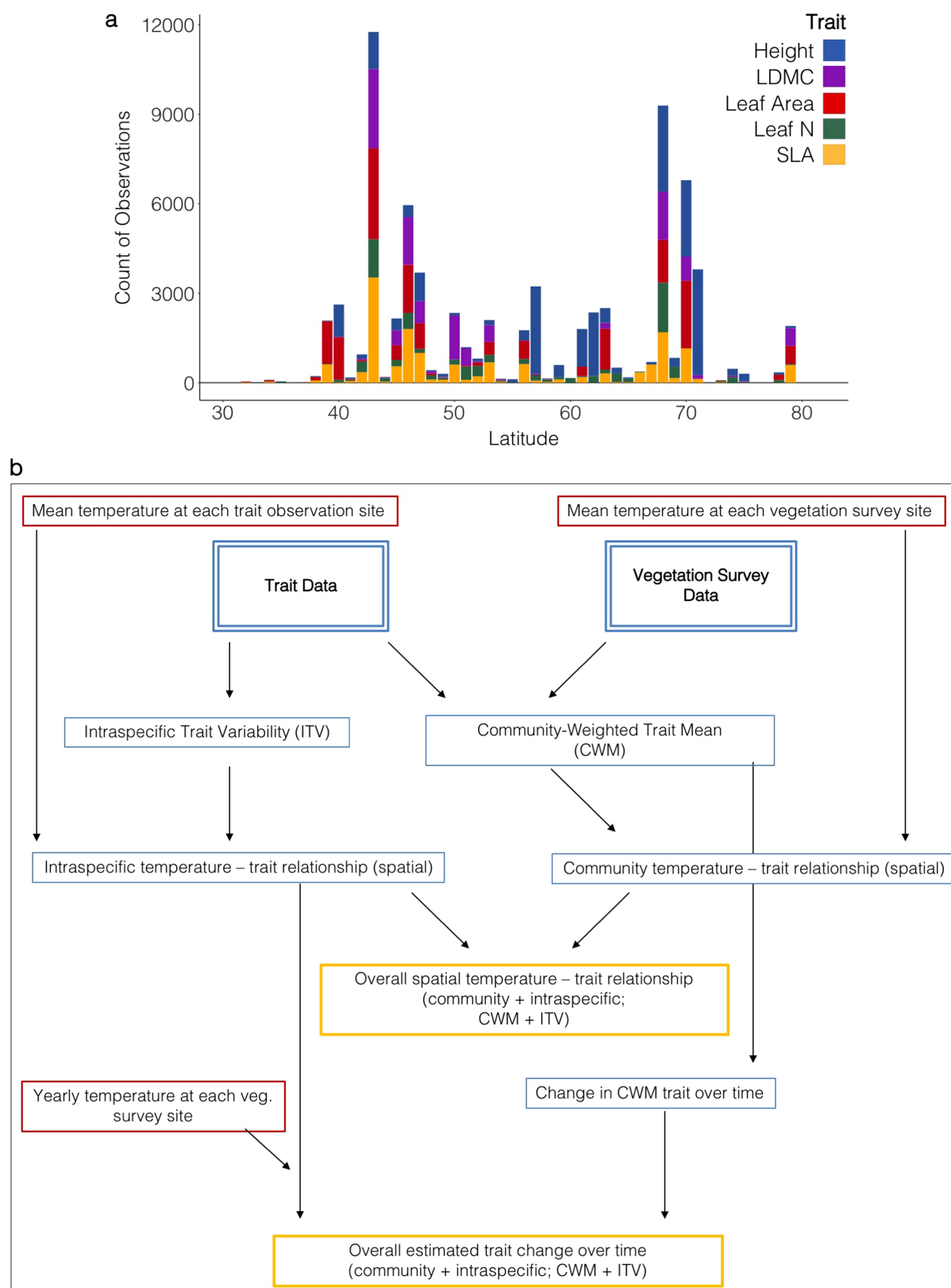
Data availability

Trait data. Data compiled through the Tundra Trait Team are publicly accessible⁵⁰. The public TTT database includes traits not considered in this study as well as tundra species that do not occur in our vegetation survey plots, for a total of nearly 92,000 trait observations on 978 species. Additional trait data from the TRY trait database can be requested at <https://www.try-db.org/>.

Composition data. Most sites and years of the vegetation survey data included in this study are available in the Polar Data Catalogue (ID 10786_1so). Much of the individual site-level data has additionally been made available in the BioTIME database⁶⁰ (<https://synergy.st-andrews.ac.uk/biotime/biotime-database/>).

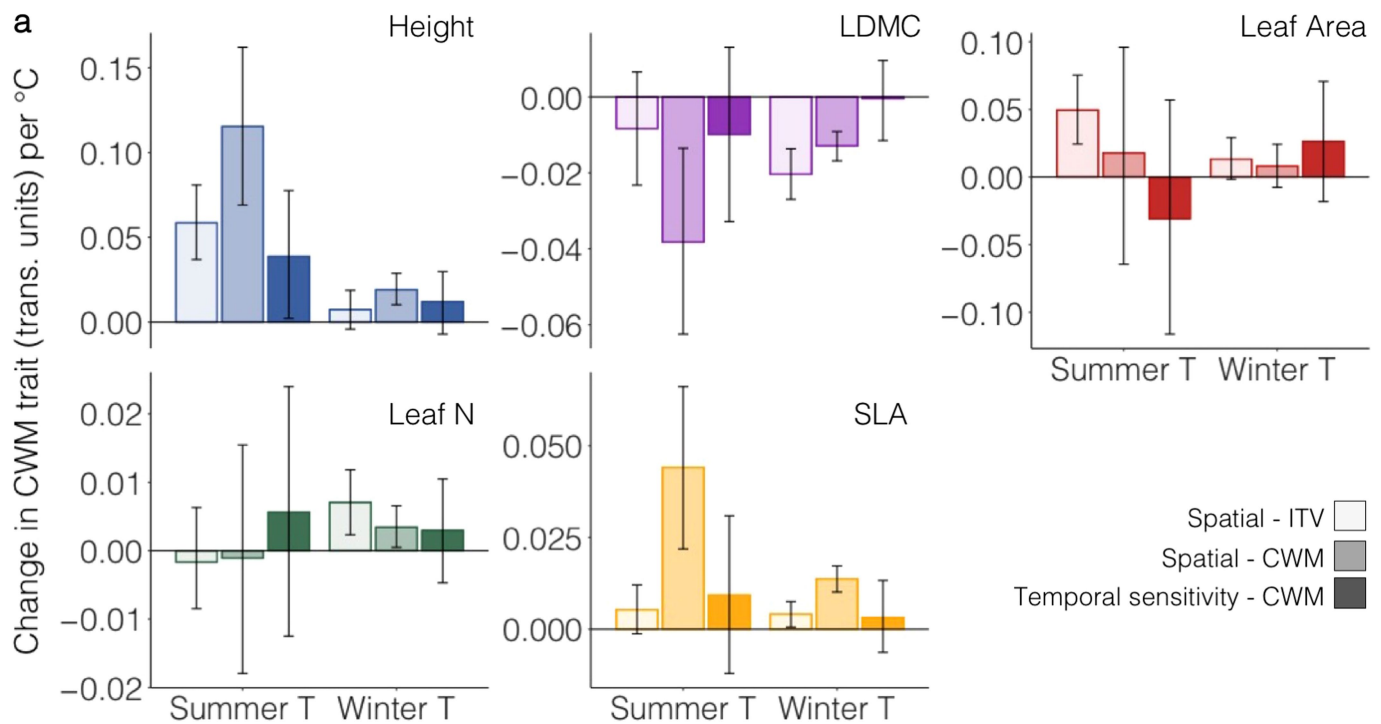
- Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
- Blok, D. et al. The cooling capacity of mosses: controls on water and energy fluxes in a Siberian tundra site. *Ecosystems* **14**, 1055–1065 (2011).
- Soudzilovskaia, N. A., van Bodegom, P. M. & Cornelissen, J. H. C. Dominant bryophyte control over high-latitude soil temperature fluctuations predicted by heat transfer traits, field moisture regime and laws of thermal insulation. *Funct. Ecol.* **27**, 1442–1454 (2013).
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, J. L. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
- Myers-Smith, I. H. et al. Climate sensitivity of shrub growth across the tundra biome. *Nat. Clim. Change* **5**, 887–891 (2015).

46. Willmott, C. J. & Robeson, S. M. Climatologically aided interpolation (CAI) of terrestrial air temperature. *Int. J. Climatol.* **15**, 221–229 (1995).
47. Serna Weiland, F. C., Vrugt, J. A., van Beek, R. (L.) P. H., Weerts, A. H. & Bierkens, M. F. P. Significant uncertainty in global scale hydrological modeling from precipitation data errors. *J. Hydrol.* **529**, 1095–1115 (2015).
48. Begueria, S., Vicente Serrano, S. M., Tomás Burguera, M. & Maneta, M. Bias in the variance of gridded data sets leads to misleading conclusions about changes in climate variability. *Int. J. Climatol.* **36**, 3413–3422 (2016).
49. Kattge, J. et al. TRY—a global database of plant traits. *Glob. Change Biol.* **17**, 2905–2935 (2011).
50. Bjorkman, A. D. et al. Tundra Trait Team: a database of plant traits spanning the tundra biome. *Glob. Ecol. Biogeogr.* <https://doi.org/10.1111/geb.12821> (2018).
51. Cayuela, L., Granzow-de la Cerda, Í., Albuquerque, F. S. & Golicher, D. J. taxonstand: an R package for species names standardisation in vegetation databases. *Methods Ecol. Evol.* **3**, 1078–1083 (2012).
52. Plummer, M. rjags: Bayesian graphical models using MCMC. R package version 4.6 <https://CRAN.R-project.org/package=rjags> (2016).
53. Stan Development Team. RStan: the R interface to Stan. R package version 2.14.1 <http://mc-stan.org/> (2016).
54. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
55. Messier, J., McGill, B. J. & Lechowicz, M. J. How do traits vary across ecological scales? A case for trait-based ecology. *Ecol. Lett.* **13**, 838–848 (2010).
56. Violle, C. et al. The return of the variance: intraspecific variability in community ecology. *Trends Ecol. Evol.* **27**, 244–252 (2012).
57. Bintanja, R. & Selten, F. M. Future increases in Arctic precipitation linked to local evaporation and sea-ice retreat. *Nature* **509**, 479–482 (2014).
58. AMAP. *Snow, Water, Ice and Permafrost in the Arctic (SWIPA) 2017*. <https://www.amap.no> (Arctic Monitoring and Assessment Programme, 2017).
59. Oksanen, J., Blanchet, F., Kindt, R. & Legendre, P. vegan: Community Ecology Package. R package version 2.4.6 <https://CRAN.R-project.org/package=vegan> (2011).
60. Dornelas, M. et al. BioTIME: A database of biodiversity time series for the Anthropocene. *Glob. Ecol. Biogeogr.* **27**, 760–786 (2018).
61. Chapin, F. S. III, BretHarte, M. S., Hobbie, S. E. & Zhong, H. L. Plant functional types as predictors of transient responses of arctic vegetation to global change. *J. Veg. Sci.* **7**, 347–358 (1996).
62. Weiher, E. et al. Challenging Theophrastus: a common core list of plant traits for functional ecology. *J. Veg. Sci.* **10**, 609–620 (1999).
63. Violle, C. et al. Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).
64. Hudson, J. M. G. & Henry, G. H. R. Increased plant biomass in a high Arctic heath community from 1981 to 2008. *Ecology* **90**, 2657–2663 (2009).
65. De Deyn, G. B., Cornelissen, J. H. C. & Bardgett, R. D. Plant functional traits and soil carbon sequestration in contrasting biomes. *Ecol. Lett.* **11**, 516–531 (2008).
66. Kunstler, G. et al. Plant functional traits have globally consistent effects on competition. *Nature* **529**, 204–207 (2016).
67. Gaudet, C. L. & Keddy, P. A. A comparative approach to predicting competitive ability from plant traits. *Nature* **334**, 242–243 (1988).
68. Westoby, M., Falster, D. S., Moldes, A. T., Vesk, P. A. & Wright, I. J. Plant ecological strategies: some leading dimensions of variation between species. *Annu. Rev. Ecol. Syst.* **33**, 125–159 (2002).
69. Moles, A. T. & Leishman, M. R. *Seedling Ecology and Evolution* (Cambridge Univ. Press, Cambridge, 2008).
70. Sturm, M. et al. Snow–shrub interactions in Arctic tundra: a hypothesis with climatic implications. *J. Clim.* **14**, 336–344 (2001).
71. Lorant, M. M., Berner, L. T., Goetz, S. J., Jin, Y. & Randerson, J. T. Vegetation controls on northern high latitude snow–albedo feedback: observations and CMIP5 model simulations. *Glob. Change Biol.* **20**, 594–606 (2014).
72. Myers-Smith, I. H. & Hik, D. S. Shrub canopies influence soil temperatures but not nutrient dynamics: an experimental test of tundra snow–shrub interactions. *Ecol. Evol.* **3**, 3683–3700 (2013).
73. DeMarco, J., Mack, M. C. & Bret-Harte, M. S. Effects of arctic shrub expansion on biophysical vs. biogeochemical drivers of litter decomposition. *Ecology* **95**, 1861–1875 (2014).
74. Enquist, B. J., Brown, J. H. & West, G. B. Allometric scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
75. Street, L. E., Shaver, G. R., Williams, M. & van Wijk, M. T. What is the relationship between changes in canopy leaf area and changes in photosynthetic CO₂ flux in arctic ecosystems? *J. Ecol.* **95**, 139–150 (2007).
76. Poorter, H. et al. Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control. *New Phytol.* **193**, 30–50 (2012).
77. Greaves, H. E. et al. Estimating aboveground biomass and leaf area of low-stature Arctic shrubs with terrestrial LiDAR. *Remote Sens. Environ.* **164**, 26–35 (2015).
78. Westoby, M. & Wright, I. J. Land-plant ecology on the basis of functional traits. *Trends Ecol. Evol.* **21**, 261–268 (2006).
79. Niinemets, Ü. A review of light interception in plant stands from leaf to canopy in different plant functional types and in species with varying shade tolerance. *Ecol. Res.* **25**, 693–714 (2010).
80. Freschet, G. T., Aerts, R. & Cornelissen, J. H. C. A plant economics spectrum of litter decomposability. *Funct. Ecol.* **26**, 56–65 (2012).
81. Manning, P. et al. Simple measures of climate, soil properties and plant traits predict national-scale grassland soil carbon stocks. *J. Appl. Ecol.* **52**, 1188–1196 (2015).
82. Iida, Y. et al. Wood density explains architectural differentiation across 145 co-occurring tropical tree species. *Funct. Ecol.* **26**, 274–282 (2012).
83. Ménard, C. B., Essery, R., Pomeroy, J., Marsh, P. & Clark, D. B. A shrub bending model to calculate the albedo of shrub-tundra. *Hydrol. Processes* **28**, 341–351 (2014).
84. Nauta, A. L. et al. Permafrost collapse after shrub removal shifts tundra ecosystem to a methane source. *Nat. Clim. Change* **5**, 67–70 (2015).
85. Hobbie, S. E. Temperature and plant species control over litter decomposition in Alaskan tundra. *Ecol. Monogr.* **66**, 503–522 (1996).
86. Weedon, J. T. et al. Global meta-analysis of wood decomposition rates: a role for trait variation among tree species? *Ecol. Lett.* **12**, 45–56 (2009).
87. Dorrepaal, E., Cornelissen, J., Aerts, R., Wallen, B. & van Logtestijn, R. Are growth forms consistent predictors of leaf litter quality and decomposability across peatlands along a latitudinal gradient? *J. Ecol.* **93**, 817–828 (2005).
88. Larsen, K. S., Michelsen, A., Jonasson, S., Beier, C. & Grogan, P. Nitrogen uptake during fall, winter and spring differs among plant functional groups in a subarctic heath ecosystem. *Ecosystems* **15**, 927–939 (2012).
89. Chapin, F. S. III, Shaver, G. R., Giblin, A. E., Nadelhoffer, K. J. & Laundre, J. A. Responses of arctic tundra to experimental and observed changes in climate. *Ecology* **76**, 694–711 (1995).
90. Reich, P. B., Walters, M. B. & Ellsworth, D. S. From tropics to tundra: global convergence in plant functioning. *Proc. Natl Acad. Sci. USA* **94**, 13730–13734 (1997).



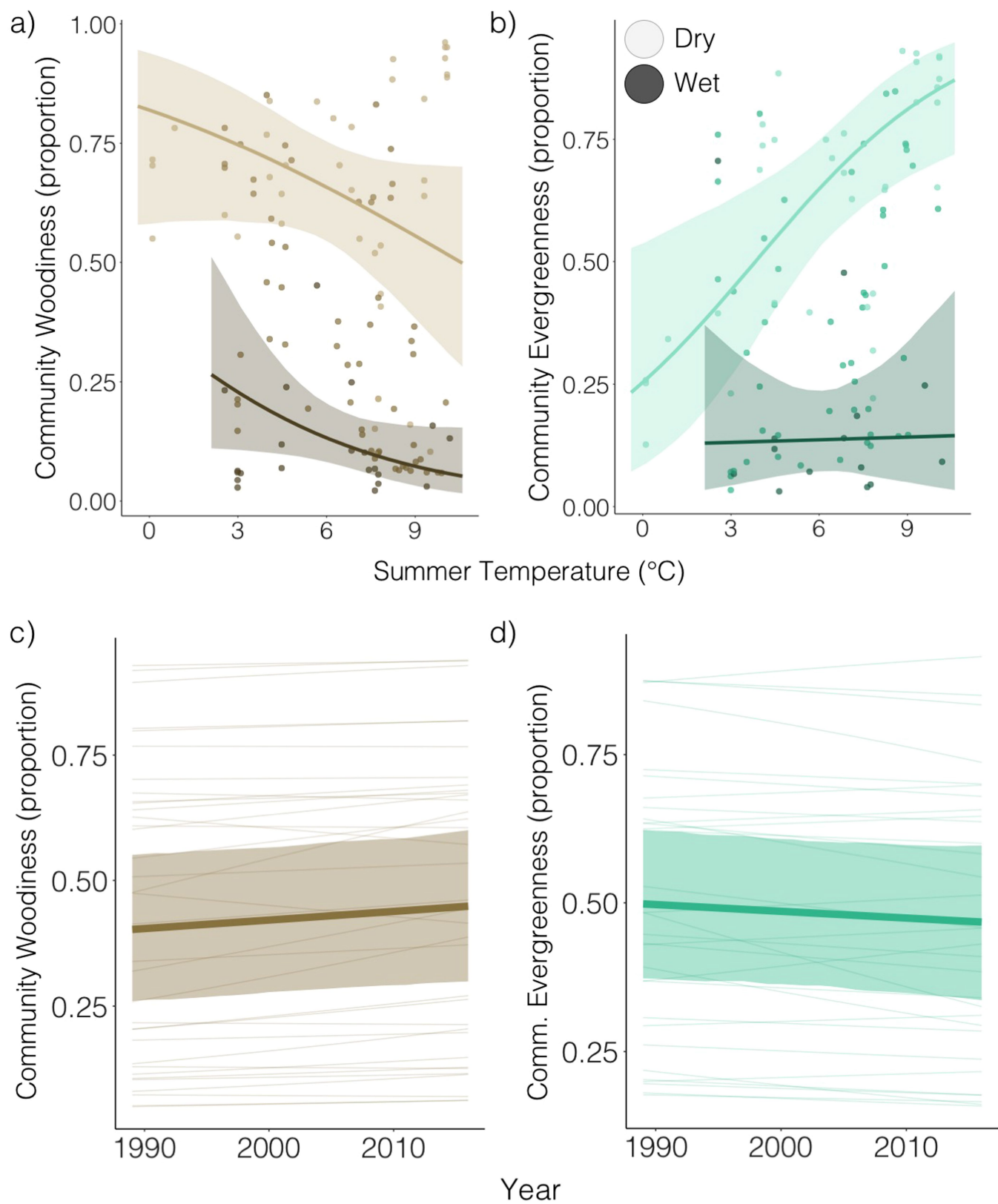
Extended Data Fig. 1 | Overview of trait data and analyses. **a**, Count of traits per latitude (rounded to the nearest degree) for all georeferenced observations in TRY and TTT that correspond to species in the vegetation survey dataset. **b**, Work flow and analyses of temperature–

trait relationships. Intraspecific temperature–trait relationships over space were used to estimate the potential contribution of ITV to overall temperature–trait relationships over space and time (CWM + ITV) as trait measurements for individual plants over time are not available.



Extended Data Fig. 2 | All temperature–trait relationships. Slope of temperature–trait relationships over space (within-species (ITV) and across communities (CWM)) and with interannual variation in temperature (community temperature sensitivity). Spatial - ITV, spatial relationship between ITV and temperature; spatial-CWM, spatial relationship between CWM and summer temperature; temporal sensitivity-CWM, temperature sensitivity of CWM (that is, correspondence between interannual variation in CWM values with interannual variation in temperature). Error bars represent 95% credible intervals on the slope estimate. We used five-year mean temperatures (temperature of the survey year and four previous years) to estimate

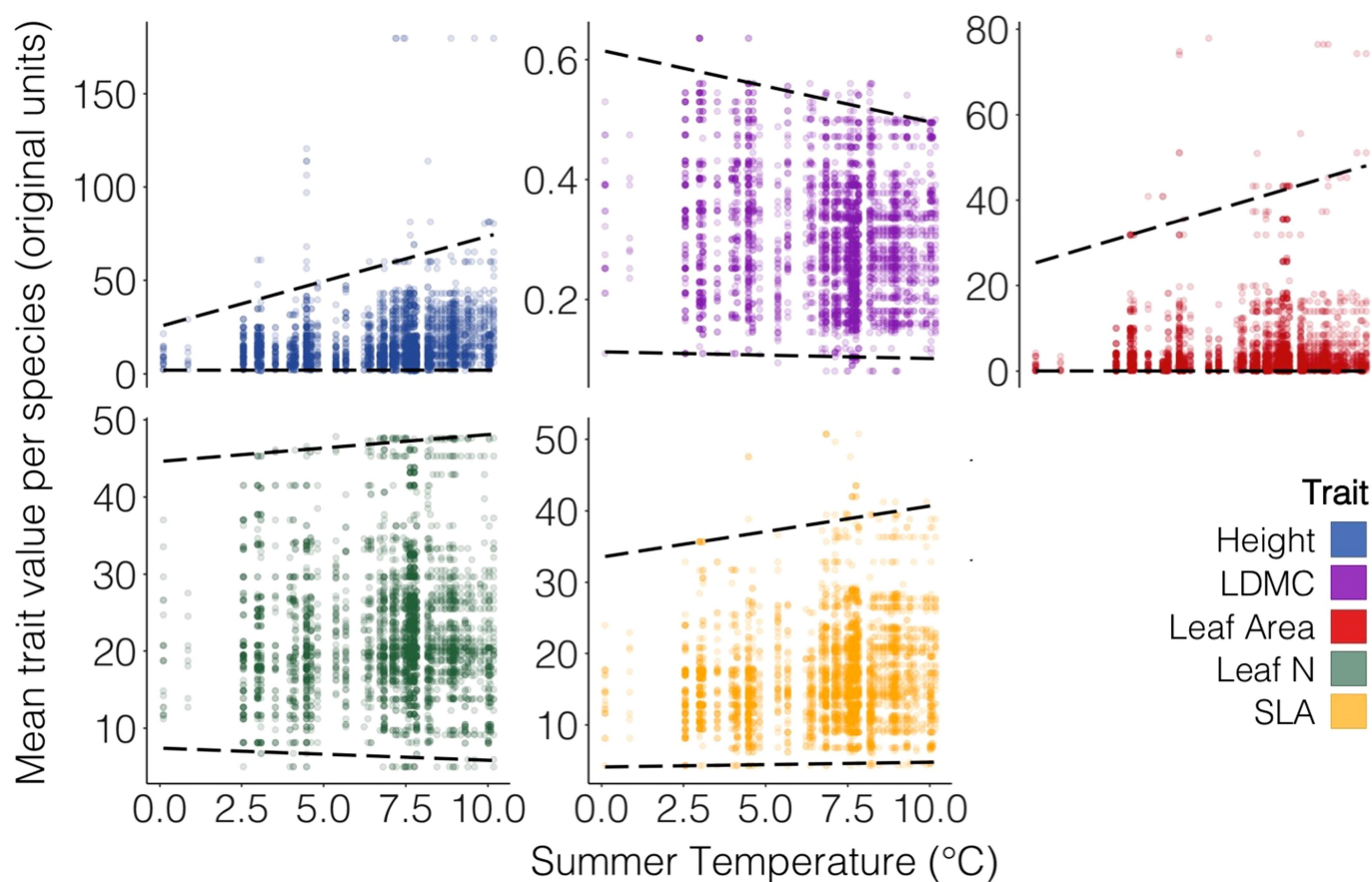
temperature sensitivity, because this interval has been shown to explain vegetation change in tundra²⁰ and alpine²⁹ plant communities. All slope estimates are in transformed units (height = $\log(\text{cm})$, LDMC = $\text{logit}(\text{g g}^{-1})$, leaf area = $\log(\text{cm}^2)$, leaf nitrogen = $\log(\text{mg g}^{-1})$, SLA = $\log(\text{mm}^2 \text{mg}^{-1})$). Community (CWM) temperature–trait relationships are estimated across all 117 sites; intraspecific temperature–trait relationships are estimated as the mean of 108 and 109 species for SLA, 80 and 86 species for plant height, 74 and 72 species for leaf nitrogen, 85 and 76 species for leaf area, and 43 and 52 species for LDMC, for summer and winter temperature, respectively (see Methods for details).



Extended Data Fig. 3 | See next page for caption.

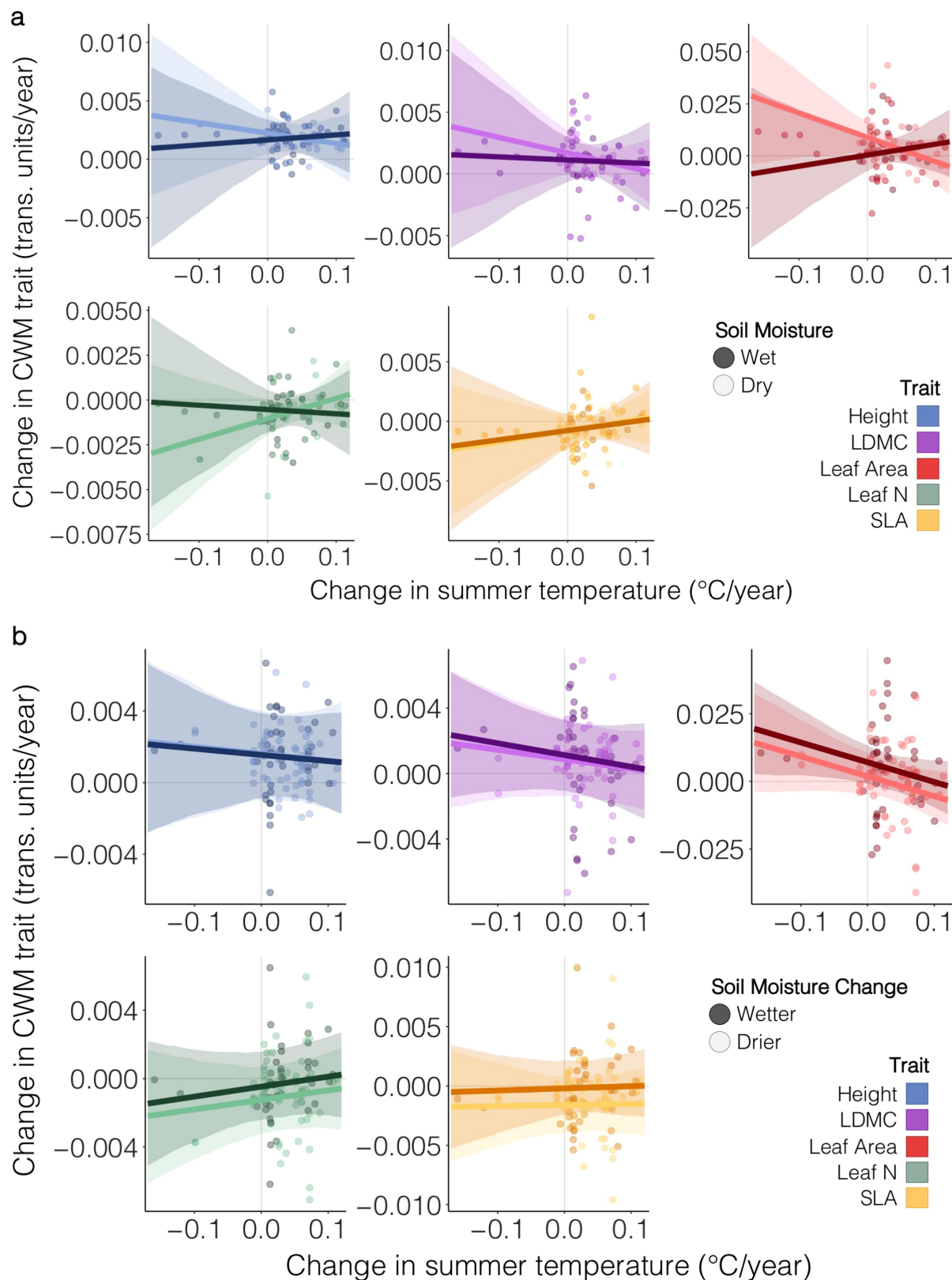
Extended Data Fig. 3 | Community woodiness and evergreenness over space and time. **a, b**, Variation in community woodiness (**a**) and evergreenness (**b**) across space with summer temperature and soil moisture. Community woodiness is the abundance-weighted proportion of woody species versus all other plant species in the community. Community evergreenness is the abundance-weighted proportion of evergreen shrubs versus all shrub species (deciduous and evergreen). The evergreen model was generated using a reduced number of sites (98 instead of 117), because some sites did not have any woody species (and it was thus not possible to calculate a proportion of evergreen species).

Both temperature and moisture were important predictors of community woodiness and evergreenness. The 95% credible interval for a temperature \times moisture interaction term overlapped zero in both models (-0.100 to 0.114 and -0.201 to 0.069 for woodiness and evergreenness, respectively). **c, d**, There was no change over time in woodiness (**c**) or evergreenness (**d**). Thin lines represent slopes per site (woodiness, $n = 117$ sites; evergreenness, $n = 98$ sites). In all panels, bold lines indicate overall model predictions and shaded ribbons designate 95% credible intervals on these model predictions.



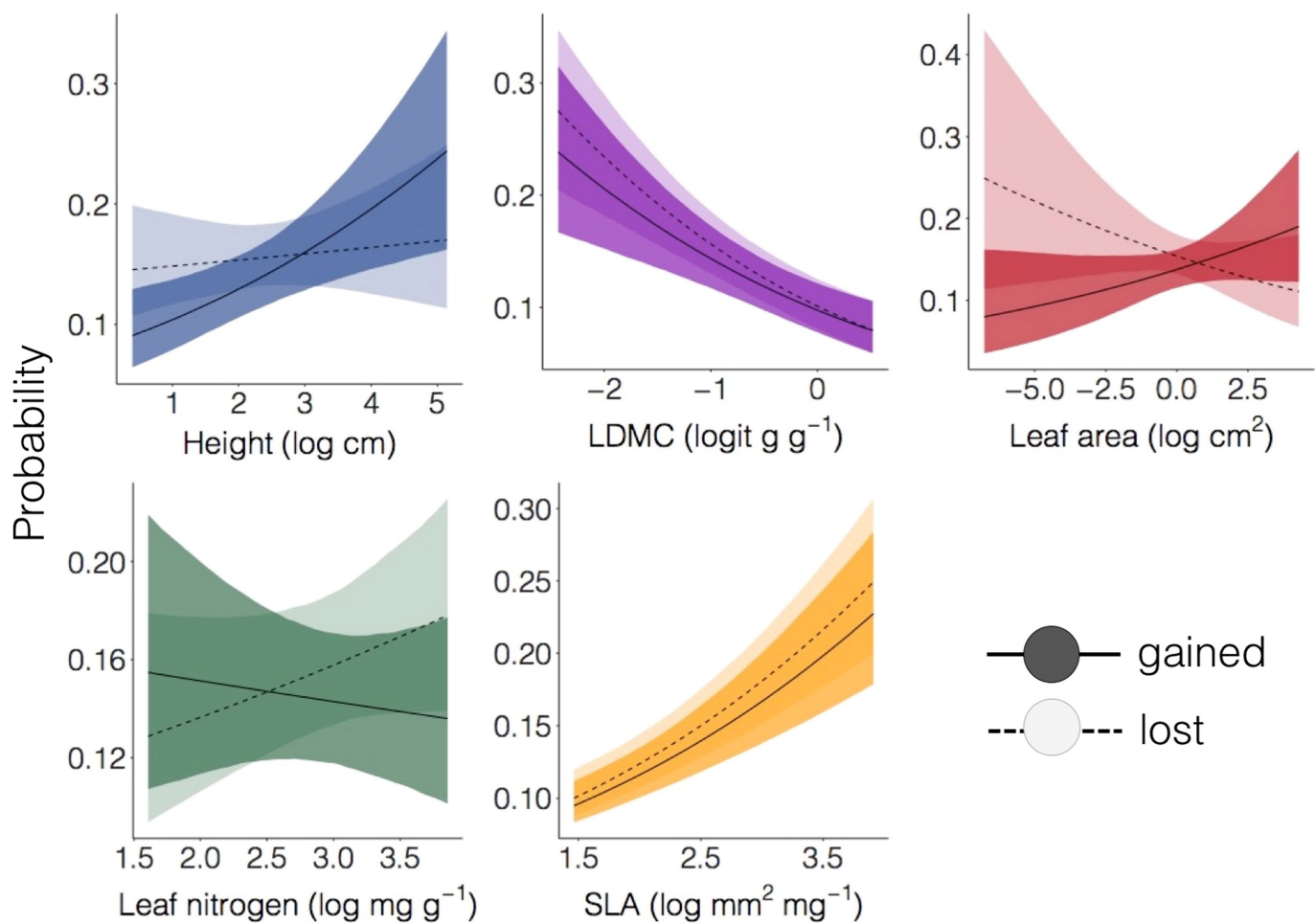
Extended Data Fig. 4 | Range in species mean values of each trait by summer temperature. Black dashed lines represent quantile regression estimates for 1% and 99% quantiles. Species mean values are estimated from intercept-only Bayesian models using the estimation technique

described in the Methods (see 'Calculation of CWM values'). Species locations are based on species in the 117 vegetation survey sites. All values are back-transformed into their original units (height (cm), LDMC (g g^{-1}), leaf area (cm^2), leaf nitrogen (mg g^{-1}), SLA ($\text{mm}^2 \text{mg}^{-1}$)).



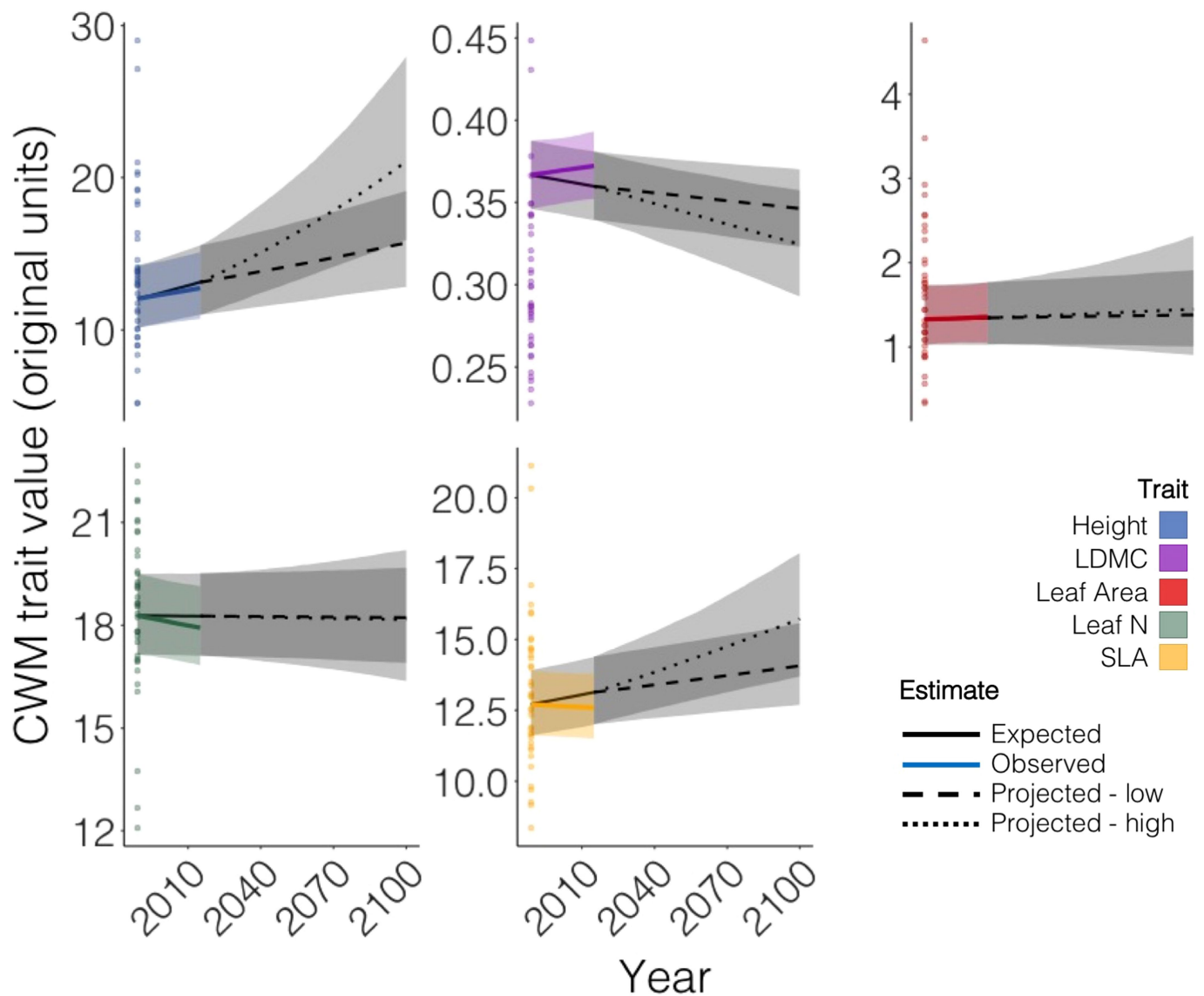
Extended Data Fig. 5 | The rate of community trait change is not related to the rate of temperature change or soil moisture for any trait. **a, b**, Rate of CWM change over time per site ($n = 117$ sites) related to temperature change and long-term mean soil moisture (**a**) or soil moisture change (**b**) at a site. Points represent mean trait change values for each site, lines represent the predicted relationship between trait change, temperature change and soil moisture or soil moisture change, and transparent ribbons are the 95% credible intervals on these predictions.

Both mean soil moisture and soil moisture change were modelled as a continuous variables, but are shown as predictions for minimum and maximum values or rates of change. Trait change estimates are in transformed units (log for height, leaf area, leaf nitrogen and SLA, and logit for LDMC). Soil moisture change was estimated from downscaled ERA-Interim data and may not accurately represent local changes in moisture availability at each site.



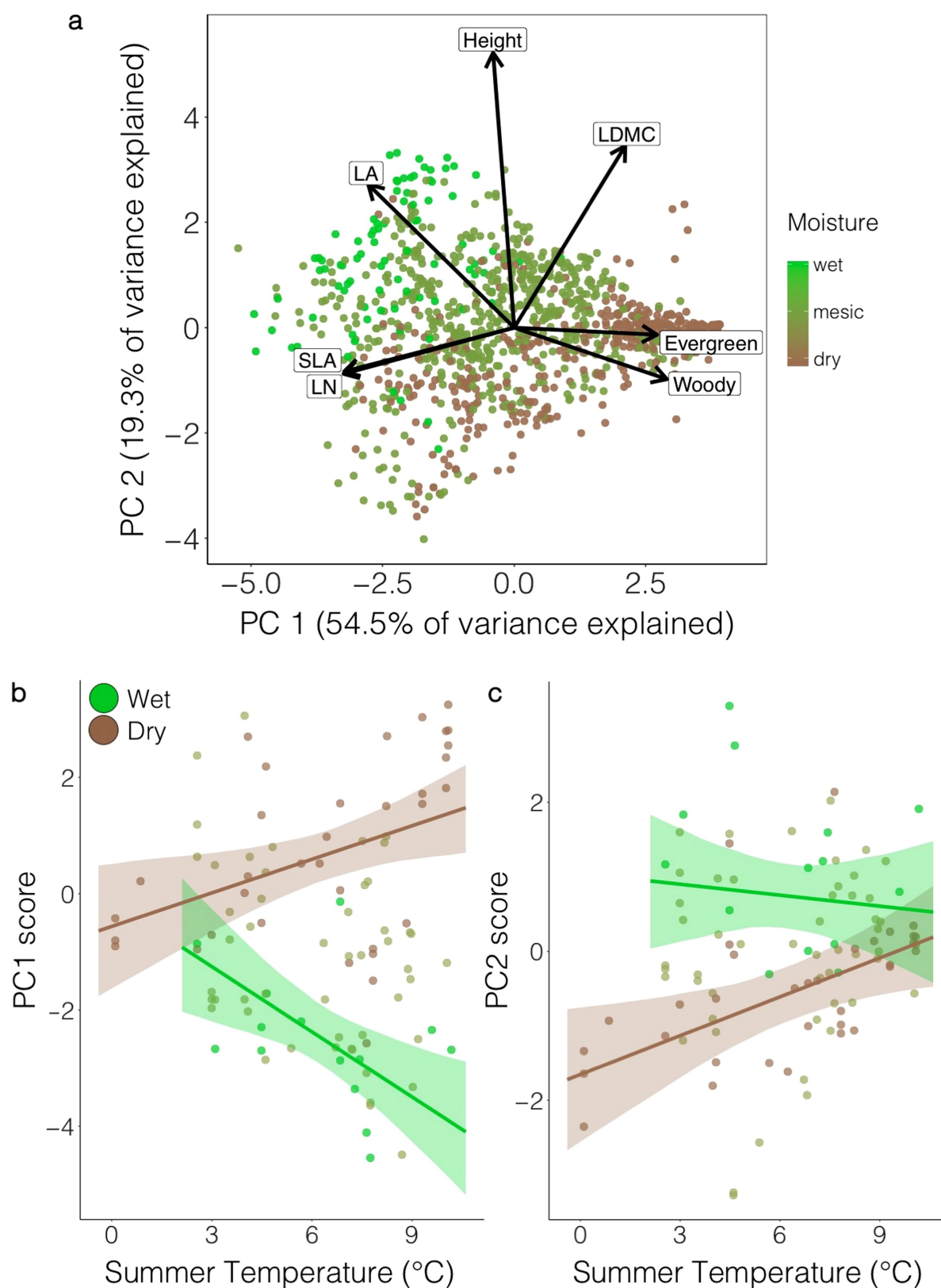
Extended Data Fig. 6 | Increasing community height is driven by the immigration of taller species, not the loss of shorter ones. Probability that a species newly arrived in a site (gained) or disappeared from a site (lost) as a function of its traits ($n = 117$ sites). Lines and ribbons represent overall model predictions and the 95% credible intervals on these

predictions, respectively. Dark ribbons and solid lines represent species gains whereas pale ribbons and dashed lines represent species losses. Only for plant height was the trait–probability relationship different for gains and losses.



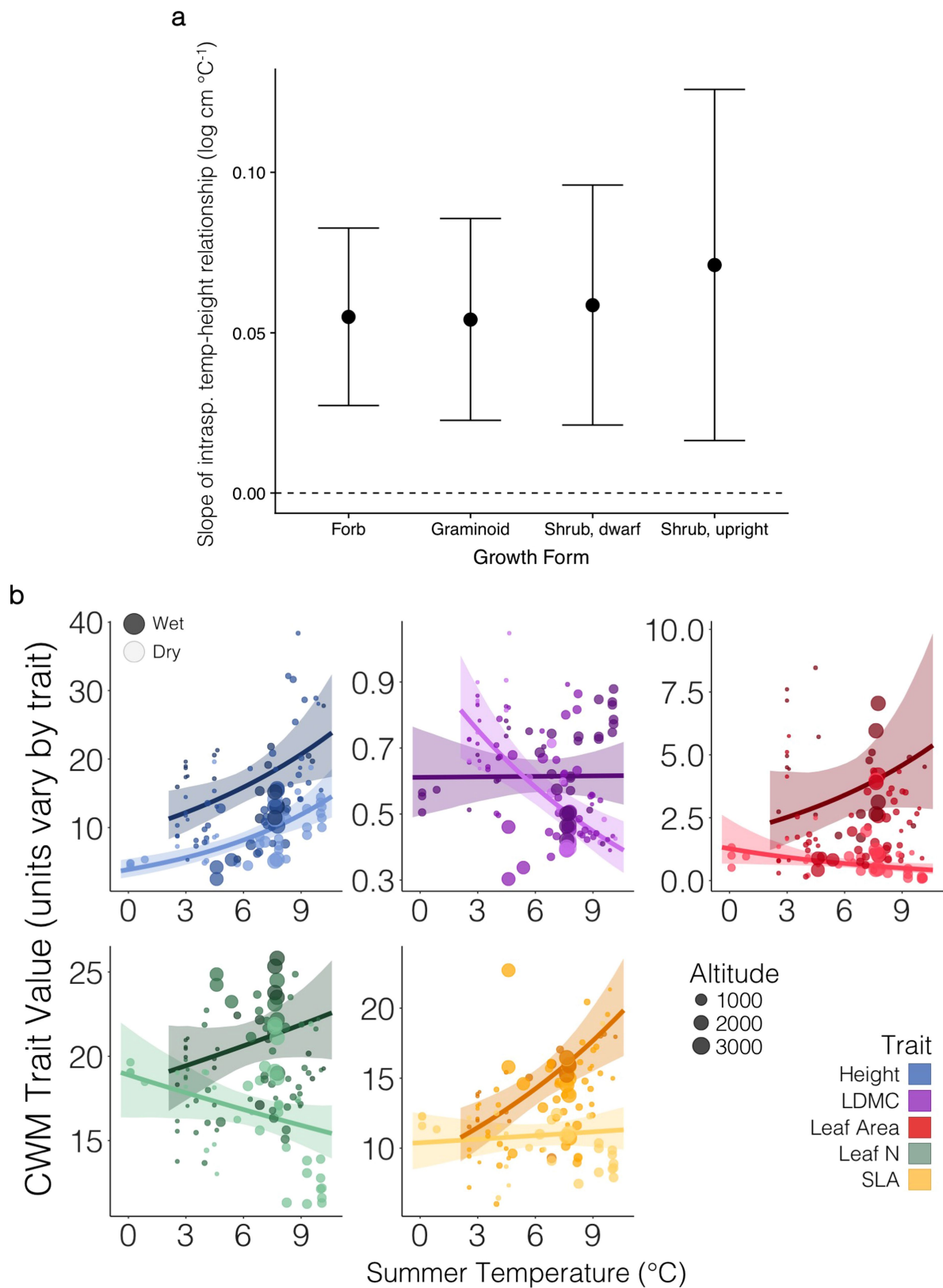
Extended Data Fig. 7 | Comparison of actual, expected and projected CWM trait change over time. Actual, expected and projected CWM trait changes are shown as solid coloured, solid black, and dashed or dotted lines, respectively. The expected trait change is calculated using the observed spatial temperature–trait relationship and the average rate of recent summer warming across all sites. Note that these projections assume no change in soil moisture conditions. The dotted and dashed black lines after 2015 show the projected trait change for the maximum

(RCP8.5) and minimum (RCP2.6) IPCC carbon emission scenarios, respectively, from the HadGEM2 AO Global Circulation Model, given the expected temperature change associated with those scenarios. Points along the left axis of each panel show the distribution of present-day CWM per site ($n=117$ sites) to better demonstrate the magnitude of projected change. Values are in original units (height (cm), LDMC (g g^{-1}), leaf area (cm^2), leaf nitrogen (mg g^{-1}) and SLA ($\text{mm}^2 \text{mg}^{-1}$)).



Extended Data Fig. 8 | Community trait co-variation is structured by temperature and moisture. **a**, PCA of plot-level community-weighted traits for seven key functional traits demonstrating how communities vary in multidimensional trait space. Trait correlations are highest between SLA and leaf nitrogen, and evergreenness and woodiness. Variation in SLA, leaf nitrogen, evergreenness and woodiness (principal component (PC)1) are orthogonal to variation in height (PC2). Variation in leaf area and LDMC are explained by both PC1 and PC2. The colour of the points indicates

the soil moisture status of each plot at the site-level. **b**, **c**, Plot scores along PC1, related to plant resource economy, vary with summer temperature, soil moisture and their interaction (**b**), whereas plot scores along PC2 vary only with soil moisture (**c**). The colour of the points indicates the soil moisture of each site. Because not all plots and sites had woody species (and thus proportion evergreen could not be calculated), this analysis was conducted on a subset of 1,098 (out of 1,520) plots at 98 (out of 117) different sites.



Extended Data Fig. 9 | Temperature–trait relationships by growth form and site elevation. a, Mean (\pm s.d.) intraspecific temperature–height relationships ($n = 80$ species) per functional group. Dwarf shrubs are defined as those shrubs that do not grow above 30 cm in height (as estimated by regional floras, such as Flora of North America, USDA or the Royal Horticultural Society) and are generally genetically limited in their ability to grow upright. There are no differences among functional groups in the magnitude of mean intraspecific temperature–height relationships.

b, Relationship between community-weighted trait values, summer temperature and soil moisture across biogeographical gradients, as in Fig. 2a. Points represent mean estimates per site ($n = 117$ sites) and are sized by the elevation of the site (larger circles indicate higher elevation). Ribbons represent the overall trait–temperature–moisture relationship (95% credible intervals on predictions at minimum and maximum soil moisture) across all sites.

Extended Data Table 1 | Ecosystem functions influenced by each of the seven plant traits

Ecosystem function	Example references
Plant Height	
Above ground biomass	Chapin et al., 1996 ⁶¹ ; Weiher et al., 1999 ⁶² ; Lavorel and Garnier, 2002 ⁶ ; Violle et al., 2007 ⁶³ ; Hudson and Henry, 2009 ⁶⁴
Carbon stock	Lavorel and Garnier, 2002 ⁶ ; De Deyn et al., 2008 ⁶⁵ ; Moles et al., 2009 ¹⁴ ; Sistla et al., 2013 ³
Light Capture	Moles et al., 2009 ¹⁴
Competition	Lavorel and Garnier, 2002 ⁶ ; Kunstler et al., 2016 ⁶⁶
Seed dispersal	Gaudet and Keddy, 1988 ⁶⁷ ; Westoby et al., 2002 ⁶⁸ ; Moles et al., 2009 ¹⁴ ; Moles & Leishman 2008 ⁶⁹
Albedo	Sturm et al., 2001 ⁷⁰ ; Sturm and Douglas, 2005 ¹² ; Loranty et al., 2014 ⁷¹
Snow cover	Sturm et al., 2001 ⁷⁰ ; Myers-Smith and Hik, 2013 ⁷² ; DeMarco et al., 2014 ⁷³
Disturbance response	Lavorel and Garnier, 2002 ⁶
Max. population density	Enquist et al. 1998 ⁷⁴
Leaf Area	
Above ground biomass	Street et al., 2007 ⁷⁵ ; Poorter et al., 2012 ⁷⁶ ; Greaves et al., 2015 ⁷⁷
Albedo	Westoby and Wright, 2006 ⁷⁸
Light interception	Niinemets, 2010 ⁷⁹ ; Díaz et al., 2016 ¹⁷
Leaf water balance	Díaz et al., 2016 ¹⁷
Leaf energy balance	Díaz et al., 2016 ¹⁷
Specific Leaf Area	
Relative growth rate	Weiher et al., 1999 ⁶² ; Wright et al., 2004 ⁸ ; Reich, 2014 ³⁷
Decomposition	Lavorel and Garnier, 2002 ⁶ ; Díaz et al., 2004 ⁹ ; Cornelissen et al., 2007 ⁵ ; Cornwell et al., 2008 ¹⁰ ; Freschet et al., 2012 ⁸⁰
Leaf life span	Reich, 2014 ³⁷ ; Wright et al., 2004 ⁸ ; Diaz et al., 2004 ⁹
Leaf Nitrogen	
Decomposition	Lavorel and Garnier, 2002 ⁶ ; Cornelissen et al., 2007 ⁵ ; Cornwell et al., 2008 ¹⁰ ; Freschet et al., 2012 ⁸⁰
Primary productivity	Weiher et al., 1999 ⁶² ; Wright et al., 2004 ⁸ ; Reich, 2014 ³⁷
Soil carbon stocks	Manning et al., 2015 ⁸¹
Leaf Dry Matter Content	
Decomposition	Lavorel and Garnier, 2002 ⁶ ; Cornelissen et al., 2007 ⁵ ; Cornwell et al., 2008 ¹⁰ ; Freschet et al., 2012 ⁸⁰
Soil carbon stocks	Manning et al., 2015 ⁸¹
Woodiness	
Plant architecture	Chapin et al., 1996 ⁶¹ ; Lida et al., 2012 ⁸²
Albedo	Sturm et al., 2001 ⁷⁰ ; Ménard et al., 2014 ⁸³
Thermal insulation	Blok et al., 2010 ³⁰ ; Myers-Smith and Hik, 2013 ⁷² ; Nauta et al., 2014 ⁸⁴
Decomposition	Hobbie, 1996 ⁸⁵ ; Cornelissen et al., 2007 ⁵ ; Weedon et al., 2009 ⁸⁶
Carbon storage	Hobbie, 1996 ⁸⁵ ; Myers-Smith et al., 2011 ¹¹ ; Sistla et al., 2013 ³
Evergreenness	
Decomposition	Dorrepaal et al., 2005 ⁸⁷ ; Cornelissen et al., 2007 ⁵ ; Cornwell et al., 2008 ¹⁰
Nutrient cycling	Larsen et al., 2012 ⁸⁸
Relative growth rate	Chapin et al., 1995 ⁸⁹ ; Reich et al., 1997 ⁹⁰

Data are from previous publications^{61–90}.

T cells in patients with narcolepsy target self-antigens of hypocretin neurons

Daniela Latorre^{1,2,10}, Ulf Kallweit^{3,4,10}, Eric Armentani¹, Mathilde Foglierini^{1,5}, Federico Mele¹, Antonino Cassotta^{1,2}, Sandra Jovic¹, David Jarrossay¹, Johannes Mathis³, Francesco Zellini⁶, Burkhard Becher⁷, Antonio Lanzavecchia¹, Ramin Khatami⁸, Mauro Manconi^{3,6}, Mehdi Tafti⁹, Claudio L. Bassetti^{3*} & Federica Sallusto^{1,2*}

Narcolepsy is a chronic sleep disorder caused by the loss of neurons that produce hypocretin. The close association with *HLA-DQB1*06:02*, evidence for immune dysregulation and increased incidence upon influenza vaccination together suggest that this disorder has an autoimmune origin. However, there is little evidence of autoreactive lymphocytes in patients with narcolepsy. Here we used sensitive cellular screens and detected hypocretin-specific CD4⁺ T cells in all 19 patients that we tested; T cells specific for tribbles homologue 2—another self-antigen of hypocretin neurons—were found in 8 out of 13 patients. Autoreactive CD4⁺ T cells were polyclonal, targeted multiple epitopes, were restricted primarily by HLA-DR and did not cross-react with influenza antigens. Hypocretin-specific CD8⁺ T cells were also detected in the blood and cerebrospinal fluid of several patients with narcolepsy. Autoreactive clonotypes were serially detected in the blood of the same—and even of different—patients, but not in healthy control individuals. These findings solidify the autoimmune aetiology of narcolepsy and provide a basis for rapid diagnosis and treatment of this disease.

Narcolepsy is a rare, life-long neurological disorder that affects about 0.05% of the general population and presents with excessive daytime sleepiness, cataplexy, hypnagogic hallucinations and sleep paralysis^{1,2}. Idiopathic sporadic narcolepsy, which represents over 98% of cases, is due to a selective loss of a small number of hypocretin (HCRT) neurons in the lateral hypothalamus^{3,4}. The strong genetic association with *HLA-DQB1*06:02*^{5,6}, the evidence for immune dysregulation^{7,8} and the increased disease incidence upon influenza vaccination^{9,10} suggest the possibility that the loss of HCRT neurons reflects the contribution of cellular and humoral immunological responses that manifest in genetically predisposed individuals upon triggering by environmental factors^{11–13}. However, so far the unequivocal demonstration of specific autoreactive T lymphocytes in narcolepsy is absent.

HCRT-specific memory CD4⁺ T cells in narcolepsy

To investigate the autoimmune basis of narcolepsy, we obtained peripheral blood samples from a total of 16 patients with HCRT deficiency (levels of HCRT in cerebrospinal fluid (CSF) < 110 pg ml⁻¹) and clinical diagnosis of narcolepsy with cataplexy (narcolepsy type 1, NT1), of whom 14 carried the disease-associated *HLA-DQB1*06:02* allele (Extended Data Table 1). We also obtained blood samples from three patients who lacked the *HLA-DQB1*06:02* allele, and who had a clinical diagnosis of narcolepsy with mild or no cataplexy and intermediate or normal levels of HCRT in the CSF (narcolepsy type 2, NT2). As controls, we obtained samples from 13 healthy donors who carried the *HLA-DQB1*06:02* allele. Patients were 16–53 years of age (median 32) and had different disease durations (range 1–36 years, median 7 years).

Given the likely low frequency of circulating autoreactive T cells^{13,14}, we used two different approaches to interrogate the T cell repertoire of patients with narcolepsy. First, memory CD45RA⁻CD4⁺ T cells were isolated by cell sorting to high purity (>98%), labelled with carboxyfluorescein succinimidyl ester (CFSE), and then stimulated with

autologous monocytes that were either untreated or pulsed with a pool of 15-mer peptides that spanned the entire sequence of the 131-amino-acid precursor molecule (prepro-hypocretin (hereafter defined as HCRT)) that gives rise—through proteolytic processing—to HCRT-1 and HCRT-2 neuropeptides¹⁵. Out of nine patients with NT1 who were analysed with this method, only one showed a clear response to HCRT on day 7 (patient P8), as demonstrated by the CFSE^{low} profile and the selective upregulation of the activation markers ICOS and CD25 (Fig. 1a). In other cases, few CFSE^{low}-proliferating T cells were detected in both unstimulated and HCRT-stimulated cultures, but only in the latter did a fraction of the cells express ICOS and CD25 (see patient P22, Fig. 1b). Of note, ICOS⁺CD25⁺ cells within CFSE^{low}-proliferating cells could be detected in six out of nine patients with NT1 (66%), but in none of six healthy controls (Fig. 1c, d).

As an alternative and more sensitive approach to identify autoreactive T cells, we used the T cell library method that we have previously found to be suitable for detecting rare antigen-specific and autoreactive T cells^{16,17}. Memory CD45RA⁻CD4⁺ T cells from 15 patients with NT1 and 3 patients with NT2 were initially expanded polyclonally, and then screened for their capacity to proliferate in response to autologous B cells pulsed with a HCRT peptide pool. With the exception of one patient (P24), all patients with NT1 or NT2 showed a clear—and often strong—proliferative response to HCRT, whereas there were only a few proliferating lines in 3 out of 12 healthy controls (Fig. 2a–c). The magnitude of the proliferative response of positive T cell lines of patients with NT1 or NT2 varied from 2.0×10^3 to 76.4×10^3 and was significantly higher compared to the proliferative response of the few positive T cell lines of controls (Fig. 2d). On the basis of these results, we estimated that the frequency of HCRT-reactive T cells in patients with NT1 or NT2 ranged from <1 to 89.7 (21.4 ± 26.4 (mean \pm s.d.), 10.5 (median)) or from 7.9 to 70.9 (36.1 ± 32 (mean \pm s.d.), 29.5 (median)), respectively, in 10^6 memory CD4⁺ T cells, which was significantly higher

¹Institute for Research in Biomedicine, Faculty of Biomedical Sciences, Università della Svizzera italiana, Bellinzona, Switzerland. ²Institute of Microbiology, ETH Zurich, Zurich, Switzerland.

³Department of Neurology, University Hospital, Bern, Switzerland. ⁴Institute of Immunology, University of Witten/Herdecke, Witten, Germany. ⁵Swiss Institute of Bioinformatics, Lausanne, Switzerland. ⁶Sleep and Epilepsy Center, Neurocenter of Southern Switzerland, Lugano, Switzerland. ⁷Institute of Experimental Immunology, University of Zurich, Zurich, Switzerland. ⁸Center for Sleep Research and Sleep Medicine, Clinic Barmelweid, Barmelweid, Switzerland. ⁹Department of Physiology, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland.

¹⁰These authors contributed equally: Daniela Latorre, Ulf Kallweit, Claudio L. Bassetti, Federica Sallusto. *e-mail: claudio.bassetti@insel.ch; federica.sallusto@irb.usi.ch

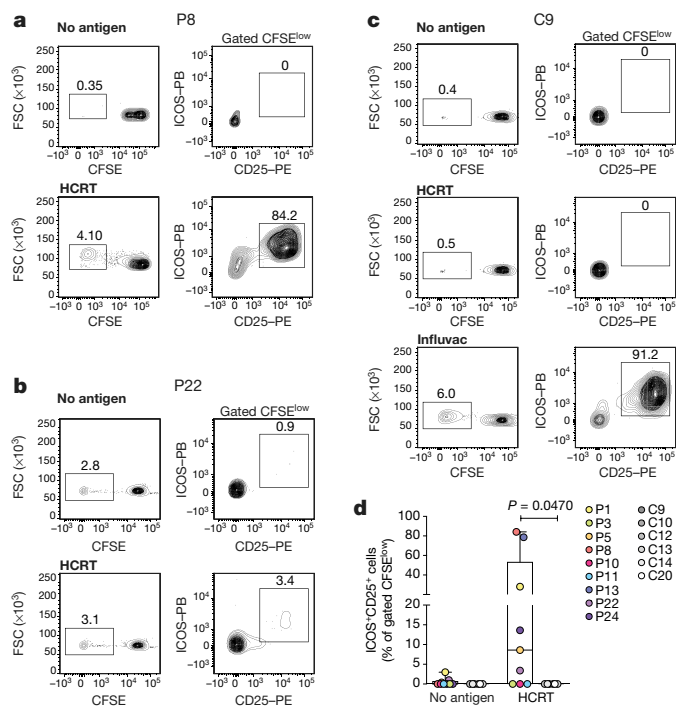


Fig. 1 | Ex vivo stimulation of memory CD4⁺ T cells from patients with narcolepsy and healthy controls. **a–c**, Memory CD4⁺ T cells from the blood of patients with narcolepsy and controls were labelled with CFSE and stimulated with autologous monocytes in the presence or absence of the HCRT peptide pool. On day seven, cells were collected and stained with anti-CD25–phycoerythrin (PE) and anti-ICOS–Pacific Blue (PB) monoclonal antibodies. CFSE profiles and dot plots of CD25 and ICOS expression of gated proliferating CFSE^{low} cells from two representative patients with NT1 (P8 (**a**) and P22 (**b**)) and one control (C9 (**c**)) are shown. Data from the stimulation of T cells from the control C9 with the influenza vaccine Influxac, used as positive control, are also shown. **d**, Pooled data from the indicated patients with NT1 ($n = 9$ biologically independent samples, coloured dots) and controls ($n = 6$ biologically independent samples, white and grey dots) are shown as the percentage of CD25⁺ICOS⁺ cells of gated proliferating CFSE^{low} cells in the absence (no antigen) or presence of the HCRT peptide pool. There were no CD25⁺ICOS⁺ cells in the pool of CFSE^{high} non-proliferating cells. Each dot represents a donor and boxes are quartile values, whiskers represent the highest and lowest values, and lines represent the median values. Data were analysed using two-tailed Mann–Whitney U -test. FSC, forward scatter.

than the frequency in controls (range <1–5.2 (1.8 ± 1.6 (mean \pm s.d.), 1.0 (median)) in 10^6 memory CD4⁺ T cells) (Fig. 2e). Of note, high frequencies of HCRT-reactive T cells were found in patients P10, P5 and P13—who were analysed 27, 6 and 2 years, respectively, after the diagnosis of NT1—and in patient P4, who was analysed 5 years after the diagnosis of NT2. Parallel screening of the same T cell libraries with seasonal influenza A antigens showed that the proliferative response and the frequency of specific memory CD4⁺ T cells were comparable in patients and controls (Extended Data Fig. 1).

Collectively, the results obtained using ex vivo antigenic stimulation and the sensitive T cell library screening method demonstrate that HCRT-specific memory CD4⁺ T cells are present in the blood of all of the patients with narcolepsy who had HCRT deficiency, even years after disease onset. HCRT-specific memory CD4⁺ T cells are also present in the patients with narcolepsy who do not carry the *HLA-DQB1*06:02* allele, and in patients with narcolepsy without HCRT deficiency.

T cells against antigens of HCRT neurons

We took advantage of the possibility of interrogating the T cell libraries with multiple antigens to investigate whether patients with narcolepsy would react to other proteins expressed by HCRT neurons. Memory

CD4⁺ T cell libraries from patients and controls were re-screened for reactivity against tribbles homologue 2 (TRIB2), an intracellular protein produced by HCRT neurons and other cell types that is targeted by antibodies in a fraction of patients with narcolepsy^{18–20}. In 8 out of 13 patients, but also in 8 out of 12 controls, we could identify several T cell lines that proliferated in response to autologous B cells pulsed with a pool of overlapping peptides spanning the TRIB2 sequence (Fig. 2f, g). Although the frequency of TRIB2-reactive memory CD4⁺ T cells was comparable, the magnitude of the proliferative response was significantly higher in patients with narcolepsy compared to controls (Fig. 2h, i). We were also able to establish T cell libraries of memory CD8⁺ T cells from 13 patients and 9 controls. Although many CD8⁺ T cell lines from both groups responded comparably to human cytomegalovirus and Epstein–Barr virus, 26 CD8⁺ T cell lines from patients—but only 2 from controls—responded to HCRT (Extended Data Fig. 2). Therefore, memory CD4⁺ and rare memory CD8⁺ T cells in patients with narcolepsy can target HCRT and other self-antigens expressed by HCRT neurons.

Characterization of autoreactive T cells

To further characterize the autoreactive T cells present in narcolepsy, we isolated 184 HCRT-specific CD4⁺ T cell clones from 9 patients and 30 TRIB2-specific CD4⁺ T cell clones from 3 patients (Supplementary Table 1), from positive cultures. Sequencing the T cell receptor β -chain variable region (*TRBV*) of all clones identified 64 HCRT-reactive and 15 TRIB2-reactive unique clonotypes. In individual patients, the response could include multiple clonotypes with no bias towards a particular V β family (Fig. 3a). When analysed for cytokine production, a panel of HCRT-reactive and TRIB2-reactive T cell clones produced IFN γ and granulocyte–macrophage colony-stimulating factor (GM-CSF) in response to antigenic stimulation (Extended Data Fig. 3a), which indicates that the memory cells are primarily of the T helper 1 (T_H1) type. HCRT- and TRIB2-autoreactive clones expressed high levels of *TBX21* and *STAT4* mRNAs as well as mRNAs encoding T_H1-associated pro-inflammatory cytokines, chemokines and chemokine receptors (Extended Data Fig. 3b).

The specificity of 57 unique HCRT-specific T cell clones and of 15 unique TRIB2-reactive T cell clones was mapped to distinct peptides that span the whole sequence of the two proteins (Fig. 3b, c). Specifically, 22 or 7 HCRT-specific T cell clones recognized peptides in HCRT-2 (amino acids 70–97) or HCRT-1 (amino acids 34–66), respectively, and 17 clones recognized peptides in the signal sequence (amino acids 1–33) or in the C-terminal region (amino acids 98–131) of HCRT¹⁵. This demonstrates that autoreactive CD4⁺ T cells target several epitopes that encompass multiple sites of HCRT-1 and HCRT-2 as well as regions that are not found in the mature proteins, with a dominance that varies among patients.

Collectively, these findings show that the response to HCRT and TRIB2 is polyclonal, polarized towards GM-CSF-producing T_H1 cells and directed against multiple epitopes in individual patients.

No T cell cross-reactivity with influenza antigens

A 6–9-fold increase in the risk of narcolepsy was reported in northern Europe after the 2009–2010 campaign of vaccination against pandemic H1N1 influenza^{9,10}, raising the possibility that the disease might be mediated by cross-reactive T cells or antibodies²¹. However, none of the HCRT- or TRIB2-specific T cell clones proliferated in response to influenza vaccine containing A/California/7/2009 H1N1 or to CA09 H1 haemagglutinin (Extended Data Fig. 4a, b). These findings do not support the notion of a molecular mimicry between epitopes on influenza antigens and HCRT, at least in this group of patients with spontaneous narcolepsy that is not associated with vaccination or infection. They are also consistent with the results of a previous study that failed to identify an immune signature induced by vaccination in its patient cohort⁷.

T cell antigen recognition and MHC restriction

The notable effect of the *HLA-DQB1*06:02* haplotype on the risk of developing narcolepsy (98% versus 25% frequency in European

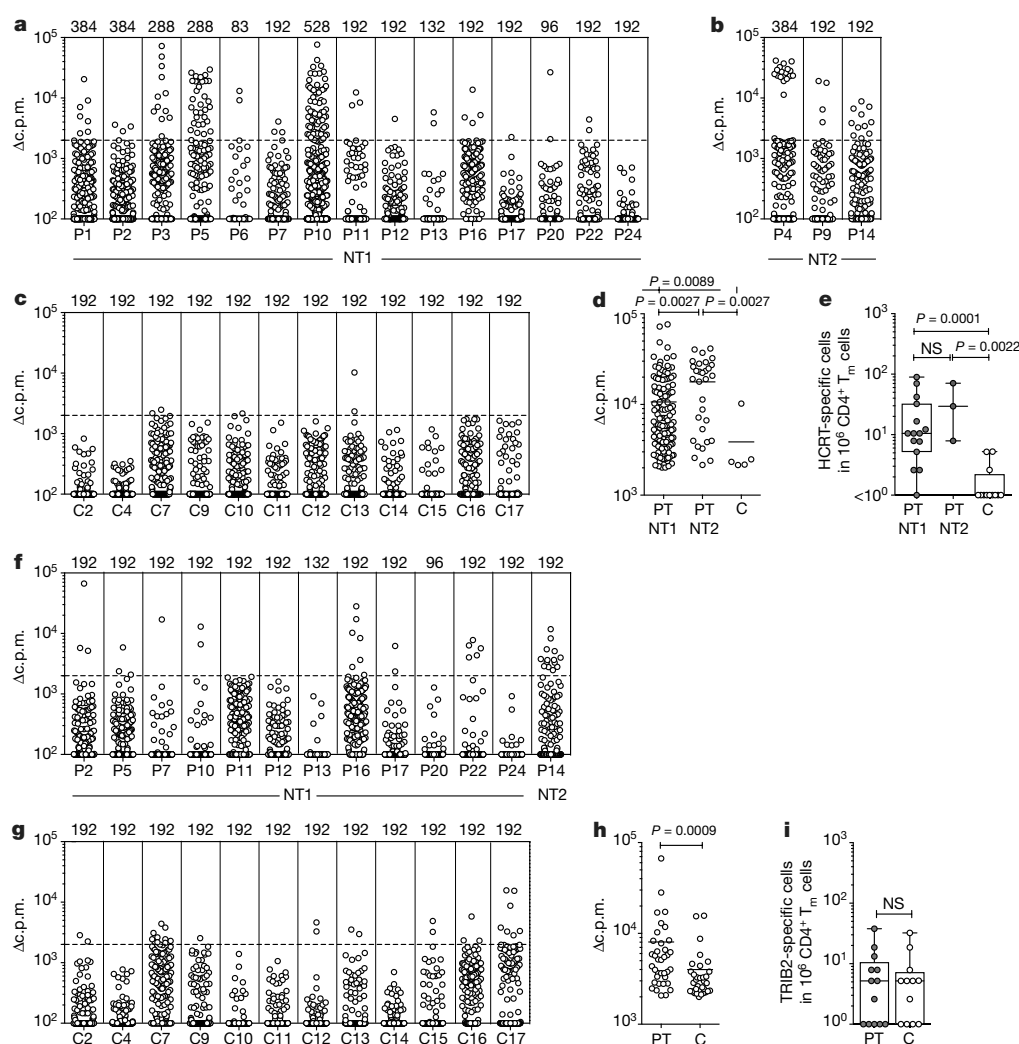


Fig. 2 | Autoreactive memory CD4⁺ T cells in patients with narcolepsy as detected using the T cell library method. **a–c, f, g.** Memory CD4⁺ T cell library screening for HCR (a–c) or TRIB2 (f, g). Memory CD45RA⁺CD4⁺ T cells from the blood of patients with NT1 or NT2 (patients labelled with P + patient number) and healthy controls (controls labelled with C + control number) were sorted and polyclonally expanded in multiple wells, each containing 2,000 cells. The number of wells ranged from 83 to 528, depending on the number of cells isolated, and is indicated on top of the graphs in a–c and in f, g. The individual T cell lines (each represented by a single dot) were screened against pools of overlapping peptides spanning the HCR (a–c) or TRIB2 (f, g) sequence, presented by irradiated autologous B cells. Proliferation was assessed on day 4 after a 16-h pulse with [³H]thymidine. Data are expressed as counts per min (c.p.m.), after subtraction of background proliferation (Δc.p.m.). The background T cell proliferation (mean ± s.d.) in the absence of antigen in patients was 1,551 ± 2,330 c.p.m. (median 859) and in controls was 958 ± 1,121 c.p.m. (median 582). Positive cultures were

defined as Δc.p.m. ≥ 2,000 (horizontal dotted line) and stimulation index (SI) ≥ 3. T cell lines with Δc.p.m. ≥ 2,000 and SI < 3 owing to ‘autologous mixed lymphocyte reaction’—which were detected in both patients and controls—were removed. **d, h.** The Δc.p.m. values of HCR-reactive lines (**d**; NT1, *n* = 140; NT2, *n* = 31; control, *n* = 5, biologically independent samples) and TRIB2-reactive lines (**h**; patients, *n* = 37; controls, *n* = 32, biologically independent samples) from patients (PT) or controls (C) are shown. **e, i.** The frequencies of HCR-specific cells (**e**) and TRIB2-specific cells (**i**) per million memory CD4⁺ T (T_m) cells in patients with narcolepsy and controls are shown (**e**, NT1, *n* = 15; NT2, *n* = 3; control, *n* = 12, biologically independent samples; **i**, patients, *n* = 13; controls, *n* = 9, biologically independent samples), calculated using the Poisson distribution. Dots represent frequency of each donor, boxes are quartile values, whiskers represent the highest and lowest values, and lines represent median values. Data were analysed using two-tailed Mann–Whitney *U*-test. NS, not significant (*P* values > 0.05).

ancestry cases and controls, respectively)^{5,6} and the association of the disease with variants in the T cell receptor-α (*TRA*) and cathepsin H (*CTSH*) genes^{22,23} suggest the importance of antigen processing and presentation to MHC-class-II-restricted T cells in the pathophysiology of narcolepsy. To investigate the basis for the association between narcolepsy and *HLA-DQB1*06:02*, we determined whether any of the clones isolated were HLA-DQ-restricted. The inhibition of peptide-induced T cell proliferation by blocking antibodies revealed that the majority of autoreactive T cell clones were HLA-DR-restricted, although a few were HLA-DQ- or HLA-DP-restricted (Fig. 3d, e and Extended Data Table 2). We then considered the possibility that the autoreactive clones might recognize HCR peptides in association with HLA-DRB1*15:01 or HLA-DRB5*01:01

molecules, the genes for which are in linkage disequilibrium with *HLA-DQB1*06:02* and were expressed in all *HLA-DQB1*06:02*-positive patients (Extended Data Table 1). A re-assessment of nine HLA-DR-restricted T cell clones demonstrated that four clones proliferated in response to a HCR peptide pool presented by autologous B cells and to a HCR peptide pool presented by an HLA-DRB1*15:01-expressing B cell line (Extended Data Fig. 4c).

To investigate the mode of antigen presentation of neuronal antigens to autoreactive T cell clones, we compared the proliferative response elicited by peptides versus soluble proteins (human TRIB2 or a mixture of human HCR-1 and HCR-2), which require processing by antigen-presenting cells (Fig. 3f). Only 1 out of 34 HCR-specific CD4⁺ T cell clones and 6 out of 15 TRIB2-specific CD4⁺ T cell clones proliferated in

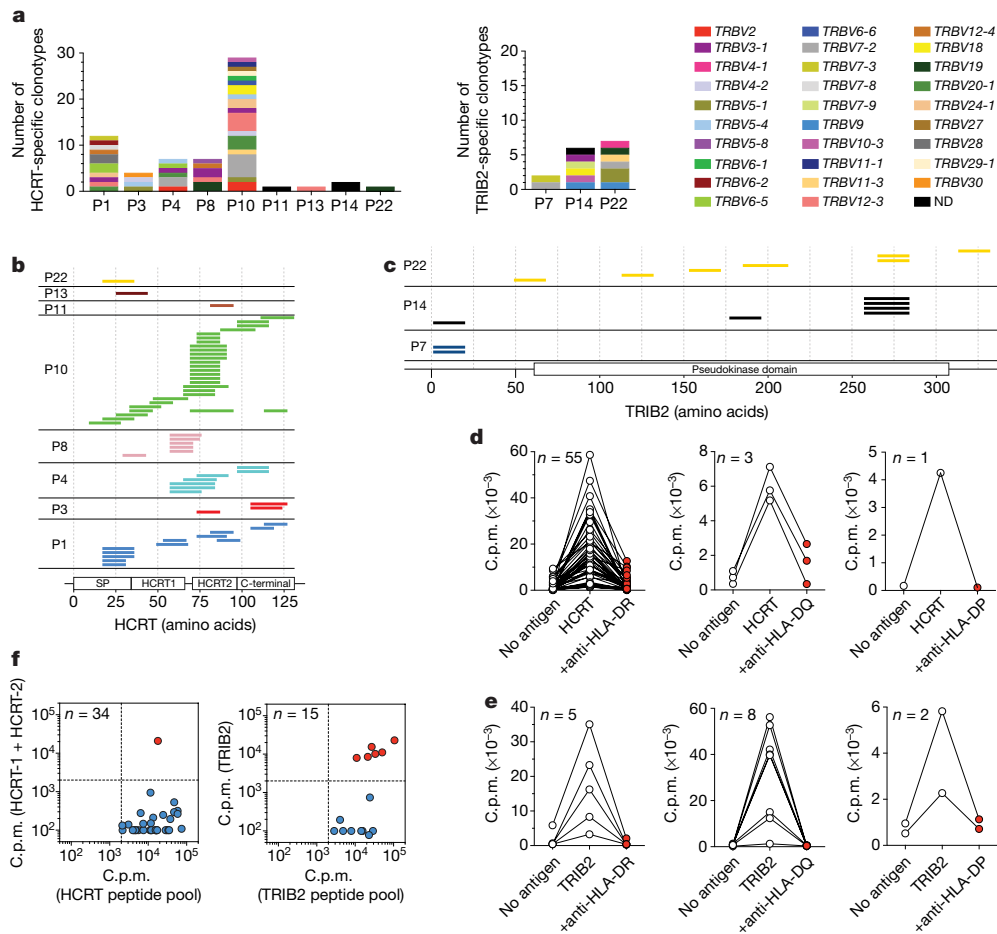


Fig. 3 | Characterization of autoreactive CD4⁺ T cell clones from patients with narcolepsy. **a**, TCR V β gene repertoire of HCRT-specific (left) or TRIB2-specific (right) CD4⁺ T cell clones isolated from the indicated patients with narcolepsy. The y axis indicates the number of autoreactive clonotypes (that is, the number of T cell clones carrying different TCR V β and complementarity-determining region 3 (CDR3) sequences). ND, not determined. **b**, **c**, Epitope mapping of HCRT-specific (**b**, $n = 57$) or TRIB2-specific (**c**, $n = 15$) CD4⁺ T cell clones from patients with narcolepsy. Epitopes were identified by screening the autoreactive CD4⁺ T cell clones against overlapping peptides that span the entire HCRT or TRIB2 protein length. Each line represents the sequence recognized by a unique clonotype and each colour indicates a patient. **d**, **e**, MHC restriction of autoreactive CD4⁺ T cell clones from patients with narcolepsy. HCRT-specific (**d**, $n = 59$) and TRIB2-specific (**e**, $n = 15$) CD4⁺ T cell clones were stimulated with autologous B cells untreated

(no antigen) or pulsed with the indicated antigens in the absence or presence of MHC-class-II blocking antibodies (+ anti-HLA-DR, + anti-HLA-DQ and + anti-HLA-DP). Proliferation was measured on day 3 after a 16-h pulse with [³H]thymidine, and is expressed as c.p.m. Each clone was tested separately with the three antibodies and HLA restriction was determined when inhibition was >80% (red dots). **f**, Type-A and type-B autoreactive CD4⁺ T cell clones from patients with narcolepsy. HCRT-specific ($n = 34$) and TRIB2-specific ($n = 15$) CD4⁺ T cell clones were stimulated with autologous B cells in the presence of antigen provided as an HCRT or TRIB2 peptide pool (x axis) or a mix of HCRT-1 and HCRT-2 or TRIB2 (y axis) protein. The HCRT-specific clones used in this assay recognized epitopes present in HCRT-1 or HCRT-2. Proliferation was measured on day 3 after a 16-h pulse with [³H]thymidine and expressed as c.p.m. Type-A and type-B clonotypes are shown in red and blue, respectively.

response to both forms of antigen (defined as type-A T cell clones), and all of the remaining clones proliferated only in response to peptides (defined as type-B T cell clones)²⁴. Comparable findings were obtained using monocytes and B cells (data not shown), which indicates that the failure to generate the correct T cell epitope is shared across different types of antigen-presenting cells. Collectively, these findings suggest that the differential recognition of proteins and peptides may be due to different complex conformations generated by peptides binding to MHC-class-II molecules in late endosomal compartments or to cell-surface or recycling MHC-class-II molecules^{24,25}.

Clonotypic analysis of blood and CSF T cells

Having obtained the TCR sequences of well-characterized autoreactive T cell clones, we next determined whether the same clonotypes could be found in a different sample from the same patient, or in different patients and in control donors. To do so, we performed high-throughput TCR V β sequencing of 0.5×10^6 memory CD4⁺ T cells directly isolated

ex vivo from blood of 18 patients (13 *HLA-DQB1*06:02*⁺ and 5 *HLA-DQB1*06:02*⁻) and 13 *HLA-DQB1*06:02*⁺ controls (Supplementary Table 2) and ranked the clonotypes according to their frequency (Fig. 4a and Extended Data Fig. 5a, b). In P10, eight TCR V β clonotypes corresponded to those of HCRT-specific T cell clones from the same patient, and two of these were among the most-frequent circulating clonotypes. In P22, three TCR V β clonotypes corresponded to those of TRIB2-specific clones, and in P3, P4, P8, P13 and P14 one or two clonotypes corresponded to sequences of autoreactive clones isolated from the same patients. Four TCR V β clonotypes associated with HCRT- or TRIB2-autoreactive T cell clones were found also in other patients (marked by an asterisk in Fig. 4a). Of note, none of the autoreactive TCR V β clonotypes was found in memory CD4⁺ T cells isolated from the blood of healthy control donors (Extended Data Fig. 5b). Overall, the TCR V β sequencing approach confirmed the presence of expanded autoreactive CD4⁺ T cell clones in the blood of patients with narcolepsy, and suggested that some autoreactive T cells may share a public TCR V β clonotype.

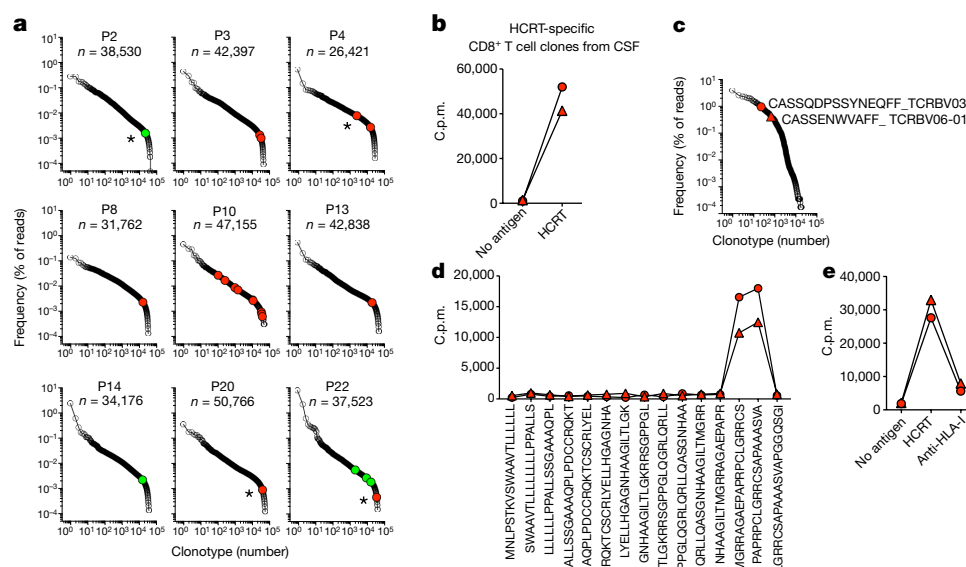


Fig. 4 | Autoreactive CD4⁺ and CD8⁺ T cell clonotypes in blood and CSF of patients with narcolepsy. a, TCR Vβ CDR3 sequences of autoreactive CD4⁺ T cell clones can be found in the blood of the same or of different patients with narcolepsy. TCR Vβ sequencing was performed on memory CD4⁺ T cells ex vivo, after sorting from peripheral blood of patients with narcolepsy. The frequency distribution of all TCR Vβ clonotypes is shown (*n* indicates total number of clonotypes). Coloured circles indicate TCR Vβ clonotypes identical to those found in HCRT-specific (red) and TRIB2-specific (green) CD4⁺ T cell clones isolated from the same patient. Asterisk indicates TCR Vβ clonotypes found in autoreactive CD4⁺ T cell clones isolated from a different patient. TCR Vβ sequencing was performed also on samples from patients P1, P5, P7, P9, P11, P12, P16, P17 and P24 and from 13 healthy controls (see Extended Data Fig. 5). In these samples, no sequences of autoreactive T cell clones

were found. **b**, Two HCRT-specific CD8⁺ T cell clones (red circle, clone 1; red triangle, clone 2) were isolated from the CSF of a patient with NT2 with recent disease onset (P14). The proliferation measured after a 16-h pulse with [³H]thymidine is expressed as c.p.m. **c**, Frequency of the identified HCRT-specific CD8⁺ T cell clones (red circle and triangle) in CSF. TCR Vβ sequencing was performed on CD8⁺ T cells sorted from in vitro-expanded CSF T cells. **d**, Epitope mapping of the identified HCRT-specific CD8⁺ T cell clones from the CSF. Epitopes were identified by screening the CD8⁺ T cell clones against overlapping peptides that span the entire length of HCRT. **e**, MHC restriction of the HCRT-specific CD8⁺ T cell clones was evaluated by measuring their proliferation against HCRT peptide pool alone or in combination with MHC-class-I blocking antibody.

We obtained a sample of CSF from 7 of our patients and—after polyclonal expansion—intrathecal CD4⁺ T cells were analysed by TCR Vβ sequencing (Supplementary Table 2), which led to the identification of 500–3,000 clonotypes in different samples. A comparison of CSF and peripheral blood showed that several clonotypes could be detected in both samples, including some that were highly represented in blood (Extended Data Fig. 6a). However, none of the autoreactive clonotypes identified in peripheral blood was also found in the CSF, possibly owing to the low frequency of the clones or to the low number of T cells analysed. We also performed TCR Vβ sequencing of CD8⁺ T cells from CSF and peripheral blood (Supplementary Table 2) and found that several clonotypes were shared between the two compartments (Extended Data Fig. 6b). Finally, we searched autoreactive T cell clones in the CSF. This approach led to the isolation of two HCRT-specific CD8⁺ T cell clones that carried different TCRs from one patient with NT2 with recent disease onset (P14). These clones recognized an epitope within the HCRT (in the region of amino acids 97–124) and were MHC-class-I-restricted, as shown by antibody inhibition experiments (Fig. 4b–e). The presence of HCRT-reactive CD8⁺ T cells in patients with normal levels of HCRT in CSF and a lack of definite cataplexy may indicate an ongoing destruction, and would be consistent with the notion that a loss of more than 80% of HCRT neurons is necessary for the development of full-blown narcolepsy with cataplexy⁴.

Discussion

The findings of this study demonstrate the existence, in patients with narcolepsy, of autoreactive memory CD4⁺ and—in some cases—CD8⁺ T cells that target self-antigens expressed by neurons that produce HCRT. The overall low frequency of autoreactive T cells may be due to the temporal gap between the onset of symptoms, the diagnosis of narcolepsy and the immunological analysis²⁶. Autoreactive CD4⁺ and CD8⁺ T cells were also found in the few cases of NT2 that we ana-

lysed; NT2 represents a less severe condition that, in some cases, can progress to full-blown NT1 with cataplexy and HCRT deficiency^{27–29}. In this context, it is of note that patient P4—who, at the time blood was drawn, had been diagnosed with NT2—recently developed cataplexy, thus meeting the clinical criteria for NT1. The previous finding of relatively high levels of CD4⁺ and CD8⁺ T cells against HCRT in this patient would be consistent with an autoimmune attack that has not (yet) led to a complete loss of neurons that produce HCRT. Future research should test for the presence of autoreactive T cells in larger populations of patients with narcolepsy, including patients with incomplete and evolving clinical manifestations as well as patients with familial and post-infectious (or post-vaccination) narcolepsy, and in populations of patients with other forms of central hypersomnolence disorders.

The finding of autoreactive CD4⁺ and CD8⁺ T cells in narcolepsy raises questions as to their possible pathogenic role. CD8⁺ T cells have the potential to directly kill HCRT neurons, as demonstrated in transgenic mice that express haemagglutinin in HCRT neurons, in which the transfer of cytotoxic haemagglutinin-specific CD8⁺ T cells led to selective neuronal destruction, sleep attack and cataplexy¹². Of note, an extensive hypothalamic CD8⁺ T cell infiltrate was reported in a patient with concomitant Ma2 antibody encephalitis, four months after the onset of symptoms of narcolepsy¹¹. By contrast, autoreactive CD4⁺ T cells may have an indirect effect that promotes the generation of pathogenic CD8⁺ T cells or autoantibodies, as hypothalamic neurons do not express constitutively MHC-class-II molecules. By producing high levels of IFNγ and GM-CSF, autoreactive CD4⁺ T cells may also promote local inflammation and loss of integrity of the blood–brain barrier, triggering the influx of effector inflammatory cells and pathogenic antibodies³⁰.

Most of the T cell clones that we isolated recognized exogenous peptides but not processed protein antigens, suggesting that self-

antigens released by dying neurons may be processed by extracellular proteases into peptides that bind to surface MHC-class-II molecules. While it is still possible that the epitopes may be generated by endogenous processing in HCRT neurons, which may be induced to express MHC-class-II molecules by IFN γ or other cytokines, it appears that in most cases professional antigen-presenting cells are unable to generate these epitopes. These findings are reminiscent of a previous report of type-B T cell clones isolated from mice with spontaneous diabetes³¹ and support the notion that extracellular processing and unconventional presentation may be a common mechanism to trigger autoreactive T cells that have escaped from central tolerance^{24,25}.

Our results do not support a molecular mimicry between HCRT and TRIB2 antigens and influenza virus, and raise questions as to the role of HLA-DQB1*06:02 in antigen presentation. The finding that only a few clones were restricted by HLA-DQ—whereas most were restricted by HLA-DR (including HLA-DRB1*15:01, the gene for which is in linkage disequilibrium with *HLA-DQB1*06:02*)—suggests that the HLA-DR-restricted response may result from an epitope-spreading phenomenon or from a compartmentalization of T cells (blood versus tissue) with different restriction, as previously reported in the case of type 1 diabetes^{32–34}.

Finally, our results may have major implications in the management of narcolepsy and may indicate new options for an earlier, more reliable and less invasive diagnosis of narcolepsy and its borderland, which currently represents a major clinical challenge. Therapeutically, our findings pave the way for systematic studies of the use of immunomodulatory interventions in narcolepsy, which in small studies have occasionally—but not invariably—been shown to have a positive effect on the evolution of the disease³⁵.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0540-1>.

Received: 5 February 2018; Accepted: 10 August 2018;

Published online 19 September 2018.

- Dauvilliers, Y., Arnulf, I. & Mignot, E. Narcolepsy with cataplexy. *Lancet* **369**, 499–511 (2007).
- Scammell, T. E. Narcolepsy. *N. Engl. J. Med.* **373**, 2654–2662 (2015).
- Peyron, C. et al. A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat. Med.* **6**, 991–997 (2000).
- Thannickal, T. C. et al. Reduced number of hypocretin neurons in human narcolepsy. *Neuron* **27**, 469–474 (2000).
- Mignot, E., Hayduk, R., Black, J., Grumet, F. C. & Guilleminault, C. HLA DQB1*0602 is associated with cataplexy in 509 narcoleptic patients. *Sleep* **20**, 1012–1020 (1997).
- Tafti, M. et al. DQB1 locus alone explains most of the risk and protection in narcolepsy with cataplexy in Europe. *Sleep* **37**, 19–25 (2014).
- Hartmann, F. J. et al. High-dimensional single-cell analysis reveals the immune signature of narcolepsy. *J. Exp. Med.* **213**, 2621–2633 (2016).
- Leclercq, M. et al. Narcolepsy type 1 is associated with a systemic increase and activation of regulatory T cells and with a systemic activation of global T cells. *PLoS ONE* **12**, e0169836 (2017).
- Pertinen, M. et al. Narcolepsy as an autoimmune disease: the role of H1N1 infection and vaccination. *Lancet Neurol.* **13**, 600–613 (2014).
- Dauvilliers, Y. et al. Increased risk of narcolepsy in children and adults after pandemic H1N1 vaccination in France. *Brain* **136**, 2486–2496 (2013).
- Dauvilliers, Y. et al. Hypothalamic immunopathology in anti-Ma-associated diencephalitis with narcolepsy–cataplexy. *JAMA Neurol.* **70**, 1305–1310 (2013).
- Bernard-Valnet, R. et al. CD8 T cell-mediated killing of orexinergic neurons induces a narcolepsy-like phenotype in mice. *Proc. Natl Acad. Sci. USA* **113**, 10956–10961 (2016).
- Ramberger, M. et al. CD4⁺ T-cell reactivity to orexin/hypocretin in patients with narcolepsy type 1. *Sleep* **40**, zsw070 (2017).
- Kornum, B. R. et al. Absence of autoreactive CD4⁺ T-cells targeting HLA-DQA1*01:02/DQB1*06:02 restricted hypocretin/orexin epitopes in narcolepsy type 1 when detected by EliSpot. *J. Neuroimmunol.* **309**, 7–11 (2017).
- Sakurai, T. et al. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
- Geiger, R., Duhon, T., Lanzavecchia, A. & Sallusto, F. Human naive and memory CD4⁺ T cell repertoires specific for naturally processed antigens analyzed using libraries of amplified T cells. *J. Exp. Med.* **206**, 1525–1534 (2009).
- Sallusto, F. et al. T-cell trafficking in the central nervous system. *Immunol. Rev.* **248**, 216–227 (2012).
- Cvetkovic-Lopes, V. et al. Elevated tribbles homolog 2-specific antibody levels in narcolepsy patients. *J. Clin. Invest.* **120**, 713–719 (2010).
- Toyoda, H. et al. Anti-tribbles homolog 2 autoantibodies in Japanese patients with narcolepsy. *Sleep* **33**, 875–878 (2010).
- Kawashima, M. et al. Anti-tribbles homolog 2 (TRIB2) autoantibodies in narcolepsy are associated with recent onset of cataplexy. *Sleep* **33**, 869–874 (2010).
- Ahmed, S. S. et al. Antibodies to influenza nucleoprotein cross-react with human hypocretin receptor 2. *Sci. Transl. Med.* **7**, 294ra105 (2015).
- Hallmayer, J. et al. Narcolepsy is strongly associated with the T-cell receptor alpha locus. *Nat. Genet.* **41**, 708–711 (2009).
- Faraco, J. et al. ImmunoChip study implicates antigen presentation to T cells in narcolepsy. *PLoS Genet.* **9**, e1003270 (2013).
- Mohan, J. F. & Unanue, E. R. Unconventional recognition of peptides by T cells and the implications for autoimmunity. *Nat. Rev. Immunol.* **12**, 721–728 (2012).
- Sadegh-Nasseri, S. & Kim, A. MHC class II auto-antigen presentation is unconventional. *Front. Immunol.* **6**, 372 (2015).
- Morrish, E., King, M. A., Smith, I. E. & Shneerson, J. M. Factors associated with a delay in the diagnosis of narcolepsy. *Sleep Med.* **5**, 37–41 (2004).
- Andlauer, O. et al. Predictors of hypocretin (orexin) deficiency in narcolepsy without cataplexy. *Sleep* **35**, 1247–1255 (2012).
- Thannickal, T. C., Nienhuis, R. & Siegel, J. M. Localized loss of hypocretin (orexin) cells in narcolepsy without cataplexy. *Sleep* **32**, 993–998 (2009).
- Baumann, C. R. et al. Challenges in diagnosing narcolepsy without cataplexy: a consensus statement. *Sleep* **37**, 1035–1042 (2014).
- Iijima, N. & Iwasaki, A. Access of protective antiviral antibody to neuronal tissues requires CD4 T-cell help. *Nature* **533**, 552–556 (2016).
- Mohan, J. F., Petzold, S. J. & Unanue, E. R. Register shifting of an insulin peptide–MHC complex allows diabetogenic T cells to escape thymic deletion. *J. Exp. Med.* **208**, 2375–2383 (2011).
- Kent, S. C. et al. Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope. *Nature* **435**, 224–228 (2005).
- Pathiraja, V. et al. Proinsulin-specific, HLA-DQ8, and HLA-DQ8-transdimer-restricted CD4⁺ T cells infiltrate islets in type 1 diabetes. *Diabetes* **64**, 172–182 (2015).
- Babon, J. A. et al. Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes. *Nat. Med.* **22**, 1482–1487 (2016).
- Barateau, L., Liblau, R., Peyron, C. & Dauvilliers, Y. Narcolepsy type 1 as an autoimmune disorder: evidence, and implications for pharmacological treatment. *CNS Drugs* **31**, 821–834 (2017).

Acknowledgements We thank all patients and their families for their participation in the study. We thank A. Sette and C. Lindstrom Arlehamm (La Jolla Institute for Allergy and Immunology) for providing the human cytomegalovirus and Epstein–Barr virus peptide pools, G. Nepom and W. Kwok (University of Washington) and R. Martin and M. Sospedra (University Hospital Zurich) for providing DR2a- and DR2b-transfected B cell lines, the Servizio Tipizzazione of the Policlinico San Matteo, University of Pavia, for HLA typing, H. Hidalgo for logistical support and L. Sallusto for discussions and support. This work was supported by the European Research Council grant (no. 323183, PREDICT, to F.S.) and the Swiss National Science Foundation grants (no. 149475 and no. CRSI13_154483 to F.S.). F.S. and the Institute for Research in Biomedicine are supported by the Helmut Horten Foundation.

Reviewer information Nature thanks B. Kornum, E. Unanue and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.L. performed experiments with assistance from E.A., F.M., A.C. and S.J. U.K., J.M., F.Z., R.K., M.M. and C.L.B. recruited participants, performed clinical evaluation and collected biological samples. D.J. performed cell sorting. M.F. performed bioinformatics analysis. M.T. supervised HLA typing. F.S. and C.L.B. conceived and supervised the project. F.S., D.L., C.L.B., U.K., M.T., B.B. and A.L. wrote the manuscript. All authors provided input and reviewed the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0540-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0540-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.L.B. or F.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Study subjects. The study included 16 patients with NT1 and 3 patients with NT2, recruited from University Hospital of Bern, Hospital of Lugano (EOC) and Clinic Barmelweid, and 13 healthy donors obtained from University Hospital of Bern (of whom one was a first-degree family member of a patient with NT1) and the Swiss Blood Donation Center of Lugano.

One patient with NT1 (P10) also suffered from multiple sclerosis. All participants provided written informed consent for participation in the study. The study was approved by the Ethical committees of Bern (no. 054/15), Lugano (no. 2909) and Barmelweid (no. 205-116). Human primary cell protocols were approved by the Federal Office of Public Health (no. A000197/2 to F.S.). All healthy donors and 14 out of 19 patients were *HLA-DQB1*06:02*-positive. All patients included in the study were HLA-typed at the Policlinico San Matteo, University of Pavia (Extended Data Table 3).

Peptides and antigens. Peptides were synthesized as crude material on a small scale (1 mg) by A and A (San Diego). Peptides used in the study included 15-mers overlapping by 11 (30 peptides) or 20-mers overlapping by 12 (15 peptides), covering the entire length of the 131-amino-acid long precursor HCRT, comprising both HCRT-1 and HCRT-2 neuropeptides as well as peptides of the signal sequence and of the C-terminal sequence; 20-mers overlapping by 12 (42 peptides) covering the entire length of TRIB2 protein; 15-mers overlapping by 10 (112 peptides) covering the haemagglutinin (HA) sequence from A/California/07/2009 (H1N1); human cytomegalovirus (CMV) and Epstein-Barr virus (EBV) HLA-class I peptides (122 peptides, 46 EBV and 76 CMV), provided by A. Sette and C. Lindstrom Arlehamn (La Jolla Institute for Allergy and Immunology). The sequences of HCRT and TRIB2 peptides are reported in Supplementary Table 3. Peptides were combined into unique pools per antigen. HCRT-1 (sequence: XPLPCCRQKTCSCRLYLHGAGNHAAGILTL, modifications: X = pyroglutamic acid, disulfide bridge between 6–12, 7–14, Leu33 = C-terminal amide) and HCRT-2 (sequence: RSGPPGLQGRLLQASGNHAAGILTM, modifications: Met28 = C-terminal amide) neuropeptides were purchased from R&D Systems, rhTRIB2 from Sino Biological and seasonal influenza virus vaccine (Influvac) from Mylan.

Cell purification and sorting. Peripheral blood mononuclear cells were isolated with Ficoll-Paque Plus (GE Healthcare). Monocytes, B cells and total CD4⁺ T cells were enriched by positive selection using CD14, CD19 and CD4-coated microbeads, respectively (Miltenyi Biotec). CD19⁺ B cells, memory CD4⁺ and CD8⁺ total cells were further sorted to over 98% purity on a FACSARIA III (BD) excluding CD45RA⁺, CD25^{bright}, CD8⁺ or CD4⁺, CD14⁺ and CD56⁺ cells. The following fluorochrome-labelled mouse monoclonal antibodies were used for staining: CD4-PE-Texas Red (clone S3.5) and CD45RA-QD655 (clone MEM-56) from Thermo Fisher Scientific, CD8-APC (clone B9.11), CD14-PE-Cy5 (clone RMO52) and CD56-PE-Cy5 (clone N901) from Beckman Coulter, CD19-FITC (clone HIB19) and CD25-PE (clone M-A251) from BD Biosciences, CCR7-BV421 (clone G043H7) from BioLegend. Cells were stained on ice for 15–20 min and sorted with FACSARIA III (BD Biosciences). CSF samples (1–2 ml) were collected by lumbar puncture. Within few hours of sampling, CSF was spun down and the pellet of cells was stimulated polyclonally with 1 µg/ml PHA (Remel) in the presence of irradiated (45 Gy) allogeneic feeder cells (1 × 10⁵ per well) and IL-2 (500 IU/ml) in a 96-well plate format. On day 15, expanded T cells were stained with CD3-PE (UCHT1), CD4-PE-Texas Red (clone S3.5), CD8-APC (clone B9.11), CD19-PE-Cy7 (SJ25C1) and CD56-PE-Cy5 (clone N901) antibodies, and CD3⁺ CD4⁺ CD8⁺ CD19⁺ CD56⁺ or CD3⁺ CD8⁺ CD4⁺ CD19⁺ CD56⁺ T cells were sorted on a FACSARIA III (BD Biosciences). The small number of CSF T cells in patients with narcolepsy limits this experimental approach and may only be informative in a small subgroup of patients with recent disease onset.

Ex vivo T cell stimulation. T cells were cultured in RPMI 1640 medium supplemented with 2 mM glutamine, 1% (v/v) nonessential amino acids, 1% (v/v) sodium pyruvate, penicillin (50 U/ml), streptomycin (50 µg/ml) (all from Invitrogen) and 5% heat-inactivated human serum (Swiss Red Cross). Sorted memory CD4⁺ T cells or expanded and sorted CD4⁺ and CD8⁺ CSF T cells were labelled with CFSE and cultured at a ratio of 2:1 with irradiated autologous monocytes untreated or pulsed for 2 h with peptide pools covering the entire sequence of HCRT or TRIB2 (3 µg/ml per peptide). After 7 days, cells were stained with antibodies to CD25-PE (clone M-A251) and ICOS-Pacific Blue (clone C398.4A) from Biolegend. The list of samples analysed ex vivo is reported in Extended Data Table 4.

T cell library. T cells were cultured in RPMI 1640 medium supplemented with 2 mM glutamine, 1% (v/v) nonessential amino acids, 1% (v/v) sodium pyruvate, penicillin (50 U/ml), streptomycin (50 µg/ml) (all from Invitrogen) and 5% human serum (Swiss Red Cross). Sorted memory CD4⁺ or CD8⁺ T cells from blood were polyclonally stimulated with 1 µg/ml PHA (Remel) in the presence of irradiated (45 Gy)

allogeneic feeder cells (5.0 × 10⁴ per well) and IL-2 (500 IU/ml) in a 96-well plate format (1,000–2,000 cells per well) and T cell lines were expanded as previously described¹⁶. Autologous B cells to be used as antigen-presenting cells were obtained by expansion with CD40L according to an established protocol³⁶. Library screening was performed at day 20–25 by culturing extensively washed T cells (~2.5 × 10⁵ cells per well) with irradiated autologous B cells (2.5 × 10⁴ cells per well), untreated or pulsed with different antigens, including HCRT peptide pool (3 µg/ml per peptide), TRIB2 peptide pool (3 µg/ml per peptide), CMV + EBV HLA-class I peptide pool (0.5 µg/ml per peptide), or influenza virus vaccine (5 µg/ml). Proliferation was measured on day 4 after 16-h incubation with 1 µCi/ml [methyl-³H]thymidine (Perkin Elmer). Precursor frequencies were calculated based on numbers of negative wells, assuming a Poisson distribution, and are expressed per million cells. Stringent criteria were used to score positive T cell lines based on a cut-off value of (i) a Δ value $\geq 2 \times 10^3$ (c.p.m. with antigen and APC – c.p.m. with APC only) and (ii) a stimulation index ≥ 3 (c.p.m. with antigen and APC/c.p.m. with APC only). This threshold was chosen based upon previous observations made across multiple negative and positive samples assessed by the T cell library technique and with a variety of donors and antigens^{16,37,38}. Note that the high sensitivity of the library method can explain differences with the ex vivo analysis, especially in the case of low-frequency, low-affinity responses such as those against self-antigens. The list of samples analysed with the T cell library method is reported in Extended Data Table 4.

Isolation of autoreactive T cell clones. To isolate autoreactive T cell clones, CFSE^{low}CD25⁺ICOS⁺ T cells from ex vivo cultures or library T cell lines were sorted and cloned by limiting dilution, as previously described³⁹. The list of samples that were cloned is reported in Extended Data Table 4. T cell clones were analysed by stimulation with irradiated autologous B cells untreated or pulsed for 2 h with HCRT or TRIB2 peptide pool (3 µg/ml per peptide) or, in some experiments, with soluble HCRT-1 and HCRT-2 (10 µg/ml each) or rhTRIB2 proteins (10 µg/ml). To determine MHC restriction, the assay was performed in the absence or presence of blocking anti-MHC-class-II antibody (anti-HLA-DR, clone L243; anti-HLA-DQ, clone SPVL3; anti-HLA-DP, clone B7/21) or blocking anti-pan-MHC-class-I antibody (clone W6/32). Some autoreactive T cell clones were also tested using different types of antigen-presenting cells, MHC-class-II-negative type-II bare lymphocyte syndrome (BLS II) B cells transfected with *DRB1*15:01* (DR2b: *DRB1*15:01*, *DRA*01:01*) or *DRB5*01:01* (DR2a: *DRB5*01:01*, *DRA*01:01*) (provided by G. Nepom and W. Kwok, University of Washington). In these experiments, autoreactive T cell clones were stimulated with irradiated autologous B cells or with irradiated DR2a- and DR2b-transfected B cell lines, pulsed or not with prepro-HCRT peptide pool (3 µg/ml per peptide). Excess of antigen was eliminated by washing 3 times after 4-h pulse of antigen-presenting cells. T cell proliferation was measured on day 3 after 16-h incubation with 1 µCi/ml [methyl-³H]thymidine (Perkin Elmer). In the cross-reactivity experiments with influenza virus antigens, autoreactive T cell clones were stimulated with irradiated autologous B cells after 2–3-h pulse with HA peptide pool (3 µg/ml per peptide) or influenza virus vaccine (5 µg/ml). Epitope mapping experiments were performed by stimulation of autoreactive T cell clones with irradiated autologous B cells after 2-h pulse with single 15-mer or 20-mer overlapping peptides (3 µg/ml per peptide) covering the whole HCRT or TRIB2 protein length. In all experiments proliferation was measured on day 3 after 16-h incubation with 1 µCi/ml [methyl-³H]thymidine (Perkin Elmer). Cytokine concentrations in the 48-h culture supernatants were assessed by Luminex bead-based assay (Thermo Fisher Scientific) according to the manufacturer's instructions.

Gene expression analysis. Autoreactive T cell clones were stimulated for 2 h in vitro with plate-bound anti-CD3 (clone TR66; 1 µg/ml) and anti-CD28 (BD Pharmingen; 1 µg/ml) monoclonal antibodies. RNA was extracted with E.Z.N.A. Total RNA kit I (Omega Biotek) following the manufacturer's instructions. Treatment with DNase enzyme (Omega Biotek) was performed during the RNA extraction procedure. RNA expression levels for a total of 594 genes were analysed by NanoString technology using nCounter Immunology Panel (Human_V2) (<https://www.nanostring.com/products/gene-expression-panels/ncounter-immunology-panels>) according to the manufacturer's instructions. In brief, 8 µl of the hybridization mix, including the Reporter CodeSet, was distributed in 12 tubes containing 25 ng total RNA in 5 µl ddH₂O from each sample (6 HCRT-specific and 6 TRIB2-specific CD4⁺ T cell clones). Subsequently, 2 µl of the Capture ProbeSet was added and, after mixing samples, the hybridization reaction was run at 65 °C for 20 h. After the incubation, samples were loaded on the Nanostring cartridge lanes following the manufacturer's instructions. The assay was run on nCounter SPRINT Profiler device. Data were normalized by the nSolver Analysis software.

Amplification of TCR V β and HLA genes. Total cDNA from individual T cell clones was obtained from 10³–10⁴ cells per reaction. Reaction was carried out using HPLC-purified oligo dT(25) primers (Microsynth) and Maxima H Minus reverse transcriptase (Thermo Fisher Scientific), in a reaction mix containing 0.2% Triton, dNTPs, Ribolock RNase inhibitor (Thermo Fisher Scientific). Reactions

were run with the following conditions: 50 °C × 60 min, 55 °C × 5 min. Five micro-litres cDNA was added to a PCR mix (final volume 25 µl) containing Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs). Sequences were amplified using TCR Vβ-specific forward primer pool, as previously described⁴⁰, and Rev84 reverse primer pairing to C1–C2 β-chain constant region with the following conditions: 98 °C × 1 min; (98 °C × 10 s; 55 °C × 20 s; 72 °C × 40 s) × 35 cycles; 72 °C × 2 min. Sequence amplification was assessed through agarose gel electrophoresis; successfully amplified fragments were sequenced by Sanger method using Rev64 primer, and TCR sequence annotation was carried out by using IMGT/V-QUEST algorithm. Forward primers. TBV-A: TCAGGT GTGATYCAATTTTC; TBV-B: AGGTGTGATCCAATTTTCG; TBV-C: TGTGTCC TGGTACCAACAG; TBV-D: GTATCGACAAGACCCAGG; TBV-E: GTATCG ACAAGACCYGGG; TBV-F: CTCACCTGAATGCCCAA; TBV-G: ATGTTYT GGTAYCGTCAG; TBV-H: CCTTACTGGTACCDGCAGA; TBV-I: ACAGAG ATGGGACAAGAAG; TBV-J: GCCATGTACTGGTAYMGA; TBV-K: CCCCAT CTCTAATCACTTATAC; TBV-L: ACATCAAACCCCAACCTATAC; TBV-M: AC CAGCAGAAGTCAAGTCA; TBV-N: TGTSTACTGGTACCARCAG; TBV-O: GGGAAAGGACAGAAAGCAAAA; TBV-P: TTAATCAGTTCCCCAGCC; TBV-Q: AGATGCAGCCCAATGAAA; TBV-R: ACAGATGGGAAACGACAA; TBV-S: GTATCRACAAGAYCCAGGA. Reverse primers. TRBC-rev84: TGTGGCCT TTTGGGTGTGG; TRBC-rev64: AGATCTCTGCTTCTGATGGC. HLA genotype of *DQB1*, *DRB1* and *DRB5* loci was determined on extracted genomic DNA (prepared using the QIAamp DNA Micro Kit, Qiagen) by PCR amplification of exon 2 using allele-specific primers on genomic DNA, followed by Sanger sequencing of the PCR products in both the forward and reverse directions. Primers targeting the intron 3 of *DRB1* were used as internal positive control of the PCR reaction. PCR for *DQB1**06:02 (mix 1, forward: CCCGAGAGGATTTTCGTGT, reverse: AACTCCGCCCCGGGTCCC, 218 bp product; mix 2, forward: CGTGCGTCTGTGACCAGAT, reverse: AACTCCGCC-CGGGTCCC, 156 bp product) and *DRB1**15:01 (forward: ACGTTTCTGTGG-CAGCCTAA, reverse: TGCAGTGTGAAGCTCTCCACAA, 262 bp product) was performed as follows: 98 °C for 1 min; (98 °C for 10 s; 72 °C for 40 s) × 36 cycles; 72 °C for 2 min. PCR for *DRB5**01:01 (forward: CTTGCAGCAGGATAAGTAT-GAG, reverse: CTGTG-AAGCTCTACCAACC, 251 bp product) and *DRB1* intron 3 (forward: TGCCAAGTG-GAGACCCCAA, reverse: GCATCTTGCTCTGTGCAGAT, 782 bp product) was performed as follow: 98 °C for 1 min; (98 °C for 10 s; 63 °C for 20 s; 72 °C for 30 s) × 36 cycles; 72 °C for 2 min. All reactions were carried out using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs), with 100 ng template gDNA in 25 µl volume.

TCR Vβ deep sequencing. Ex vivo-sorted total memory CD4⁺ or CD8⁺ T cells and in vitro-expanded and sorted CD4⁺ or CD8⁺ T cells from CSF (2.5–5 × 10⁵ cells) were analysed by deep sequencing. In brief, cells were centrifuged and

washed in PBS, and genomic DNA was extracted from the pellet using QIAamp DNA Micro Kit (Qiagen), according to manufacturer's instructions. Genomic DNA quantity and purity were assessed through spectrophotometric analysis. Sequencing of TCR Vβ CDR3 was performed by Adaptive Biotechnologies using the ImmunoSEQ assay (<http://www.immunoseq.com>). In brief, following multiplex PCR reaction designed to target any CDR3 Vβ fragments; amplicons were sequenced using the Illumina HiSeq platform. Raw data consisting of all retrieved sequences of 87 nucleotides or corresponding amino acid sequences and containing the CDR3 region were exported and further processed. The assay was performed at survey level (detection sensitivity, 1 cell in 40,000). A clonotype was defined as a unique combination of a CDR3 amino acid sequence and its related V gene. Data processing was done using the productive frequency and the diversity metrics provided by ImmunoSEQ Analyzer V3.0 (<http://www.immunoseq.com>). Antigen-specific clonotypes in each donor's repertoire were defined according to identical amino acid sequence in the Vβ CDR3 and identical V gene usage. The samples analysed are listed in Supplementary Table 2 and the TCR sequences are available as a .txt file in Supplementary Information.

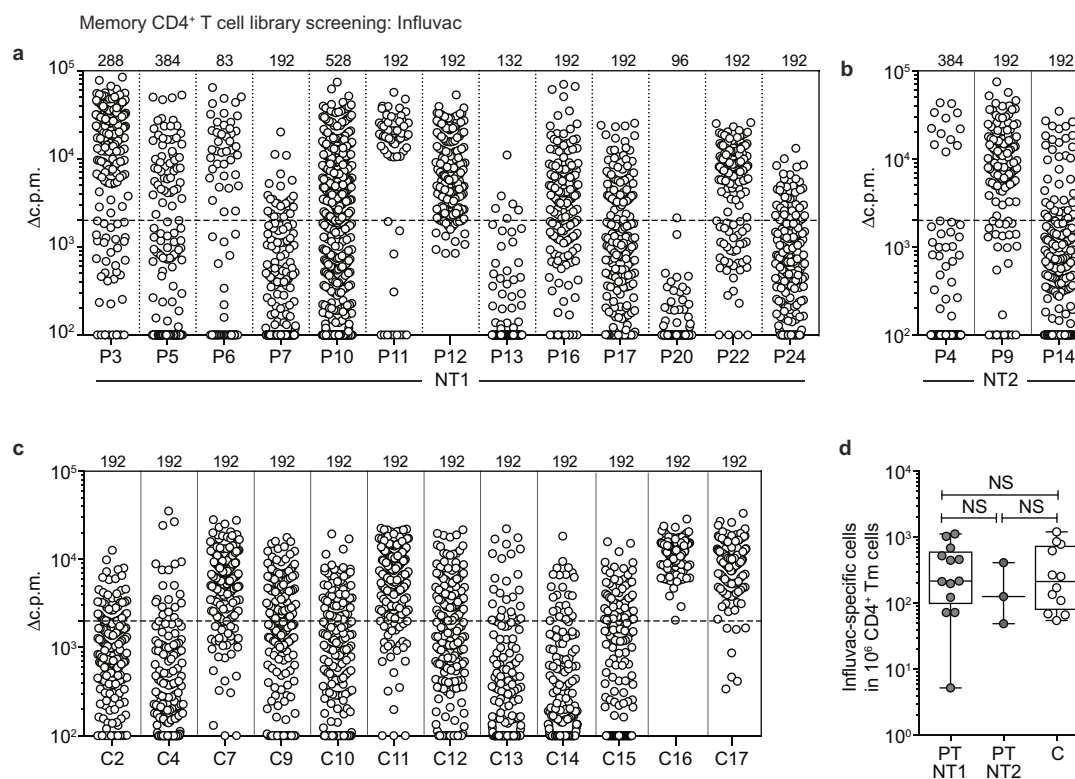
Code availability. The Java scripts used to search for antigen-specific clonotypes in donor's repertoire are available at: <https://bitbucket.org/mathildefog/narcolepsy>. This code is distributed as open source under the terms of the GNU Free Documentation License.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

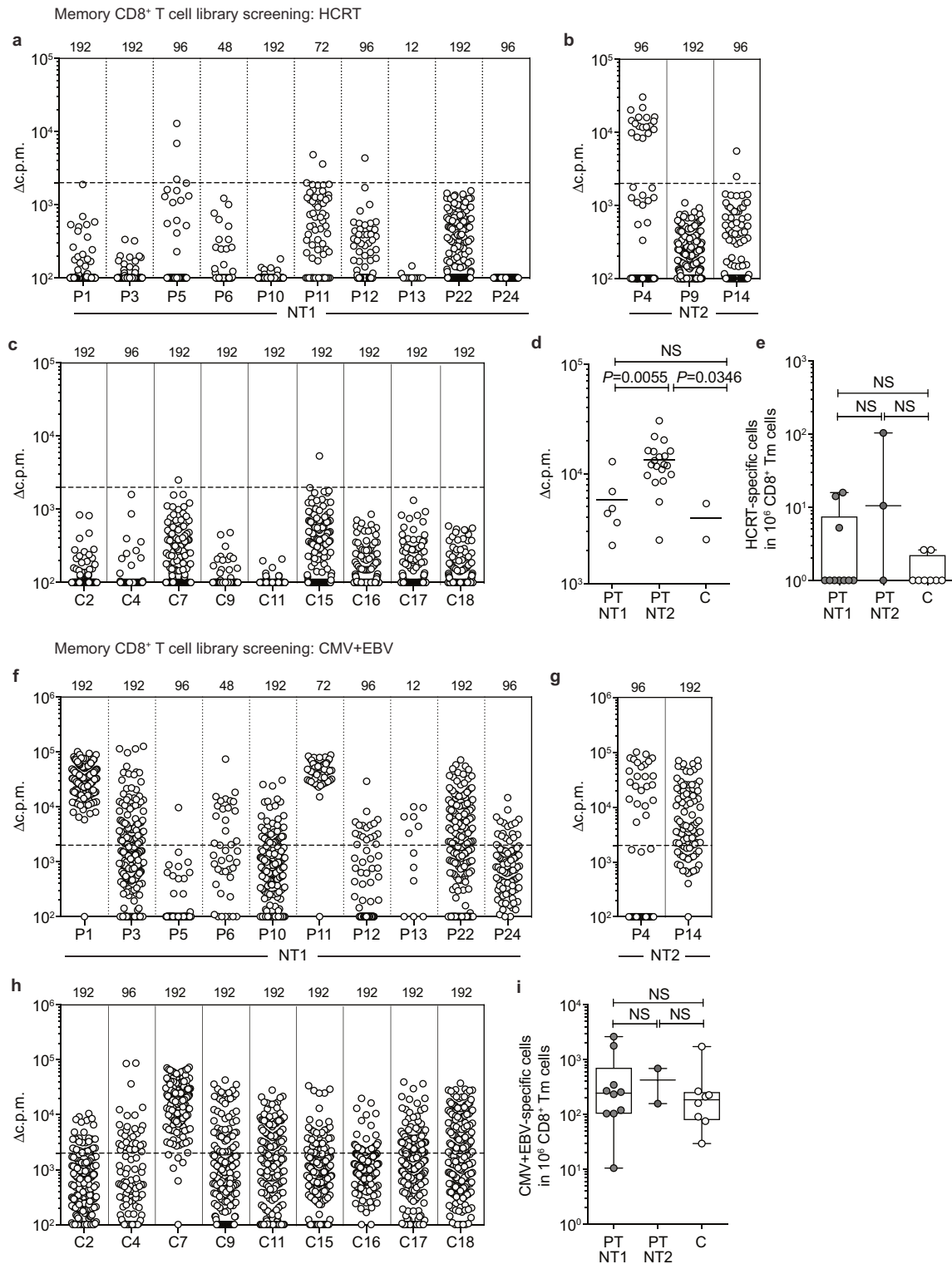
The data presented in this manuscript are included in the paper and its Supplementary Information. TCR sequences from samples listed in Supplementary Table 2 are available as a .txt file. The sequences have also been deposited in the ImmunoAccess database (<http://clients.adaptivebiotech.com/pub/Latorre-2018-nature>; <https://www.doi.org/10.21417/B73H0P>).

36. Zand, M. S. et al. A renewable source of donor cells for repetitive monitoring of T- and B-cell alloreactivity. *Am. J. Transplant.* **5**, 76–86 (2005).
37. Lindestam Arlehamn, C. S. et al. Memory T cells in latent *Mycobacterium tuberculosis* infection are directed against three antigenic islands and largely contained in a CXCR3⁺CCR6⁺ T_H1 subset. *PLoS Pathog.* **9**, e1003130 (2013).
38. Campion, S. L. et al. Proteome-wide analysis of HIV-specific naive and memory CD4⁺ T cells in unexposed blood donors. *J. Exp. Med.* **211**, 1273–1280 (2014).
39. Messi, M. et al. Memory and flexibility of cytokine gene expression as separable properties of human T_H1 and T_H2 lymphocytes. *Nat. Immunol.* **4**, 78–86 (2003).
40. Becattini, S. et al. Functional heterogeneity of human memory CD4⁺ T cell clones primed by pathogens or vaccines. *Science* **347**, 400–406 (2015).



Extended Data Fig. 1 | Screening of memory CD4⁺ T cell library from patients with NT1 or NT2 and healthy controls. a–c, Memory CD4⁺ T cell screening for Influvac. Memory CD4⁺ T cell libraries from patients with NT1 or NT2 and control donors shown in Fig. 2 were also screened for their capacity to proliferate in response to the influenza vaccine Influvac, used as a positive control. On day 4, proliferation was measured after a 16-h pulse with [³H]thymidine. The number of tested T cell lines per donor is indicated on top of the graphs and the proliferation of individual T cell lines (each line is represented by a single dot) is shown as

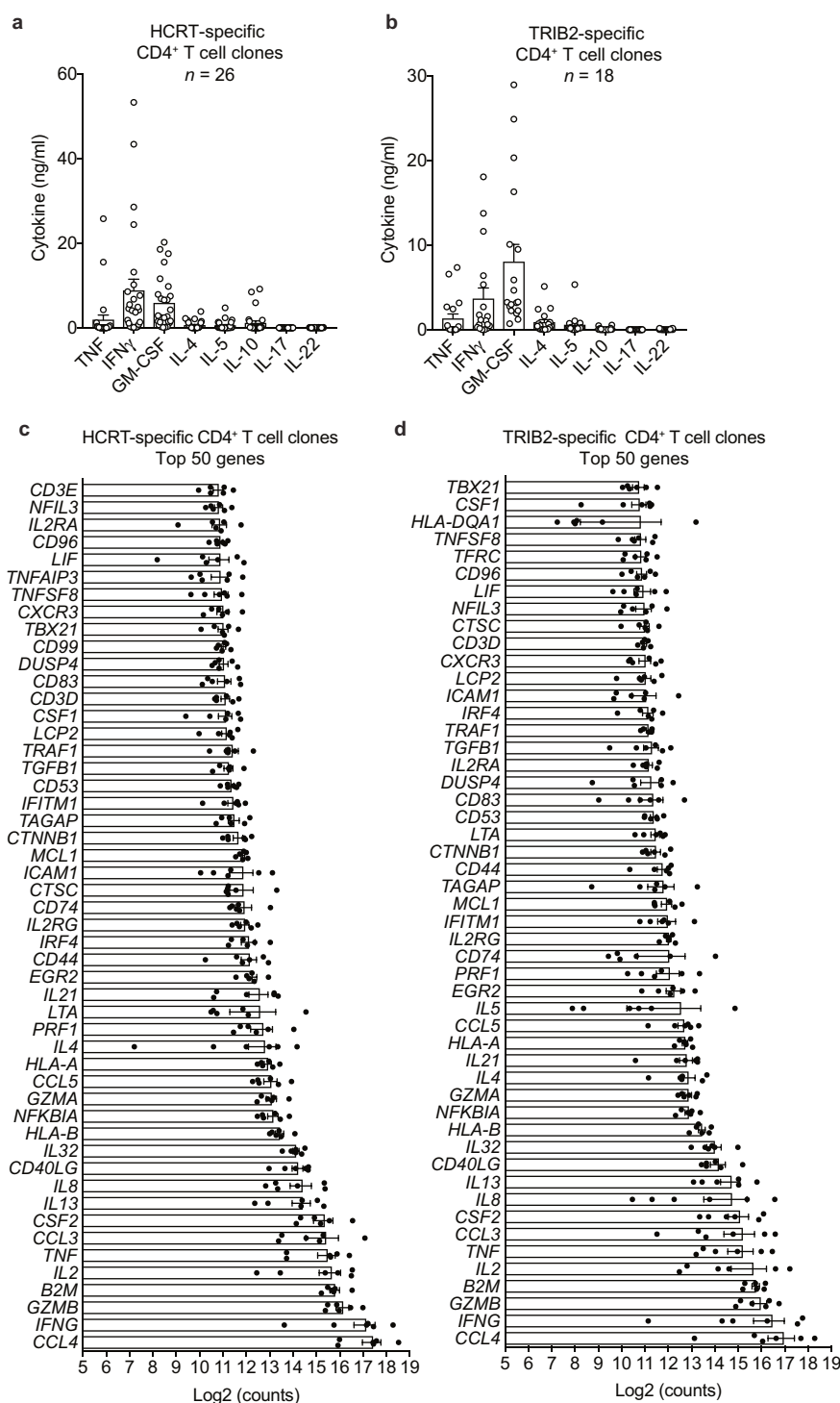
Δ c.p.m. value. Positive lines were defined as Δ c.p.m. $\geq 2,000$ (horizontal dotted line) and SI ≥ 3 . **d,** The frequency of Influvac-specific memory CD4⁺ T cells in patients with narcolepsy and controls (NT1, $n = 13$; NT2, $n = 3$; control, $n = 12$, biologically independent samples) is shown. Dots represent frequency in each donor, boxes are quartile values, whiskers represent the highest and lowest values, and lines represent the median values. Results are presented as precursor frequency per million memory CD4⁺ T cells. Data were analysed using two-tailed Mann–Whitney *U*-test. NS, not significant (P values > 0.05).



Extended Data Fig. 2 | Autoreactive memory CD8⁺ T cells in patients with NT1 or NT2 as detected using the T cell library method.

a–c, f–h. Memory CD8⁺ T cells were polyclonally expanded and screened for their capacity to proliferate in response to HCRT peptide pool (**a–c**) or CMV + EBV peptide pool (**f–h**), used as positive control, in the presence of irradiated autologous B cells. On day 4, proliferation was measured after a 16-h pulse with [³H]thymidine. The number of tested T cell lines per donor is indicated on top of the graphs and the proliferation of individual T cell lines (each represented by a single dot) is shown as Δ c.p.m. value. Positive responses were defined as Δ c.p.m. \geq 2,000 (horizontal dotted line) and SI \geq 3. **d.** The Δ c.p.m. values of HCRT-positive T cell lines (NT1, $n = 6$; NT2, $n = 20$; control, $n = 2$, biologically

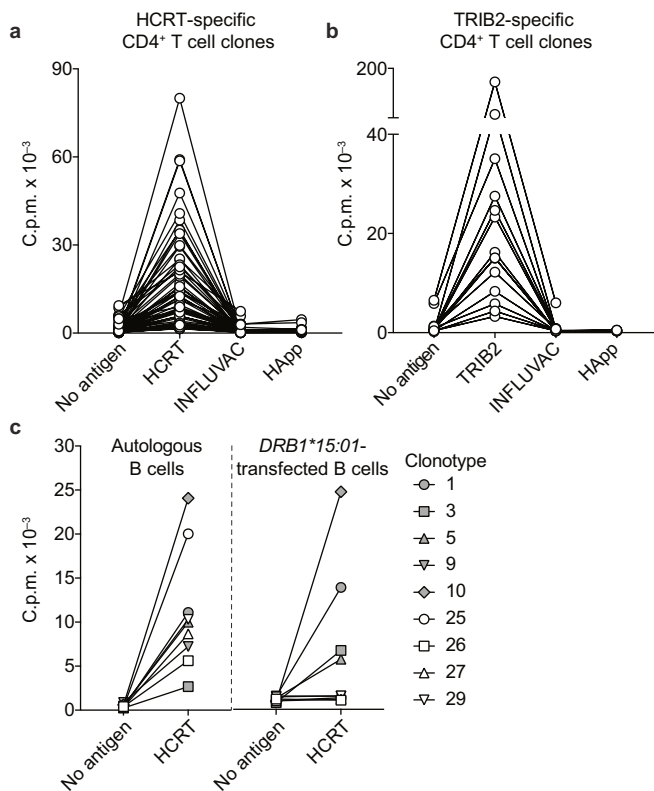
independent samples) from patients (PT) and controls (C) are shown. **e, i.** The frequency of HCRT-specific (**e**) and CMV + EBV-specific (**i**) cells per million memory CD8⁺ T cells in patients with narcolepsy and controls is shown (**e**, NT1, $n = 10$; NT2, $n = 3$; control, $n = 9$, biologically independent samples; **i**, NT1, $n = 10$; NT2, $n = 2$; control, $n = 9$, biologically independent samples). Dots represent frequency in each donor, boxes are quartile values, whiskers represent the highest and lowest values, and lines represent the median values. Results are presented as precursor frequency per million memory CD8⁺ T cells. Data were analysed using two-tailed Mann–Whitney *U*-test. NS, not significant (P values > 0.05).



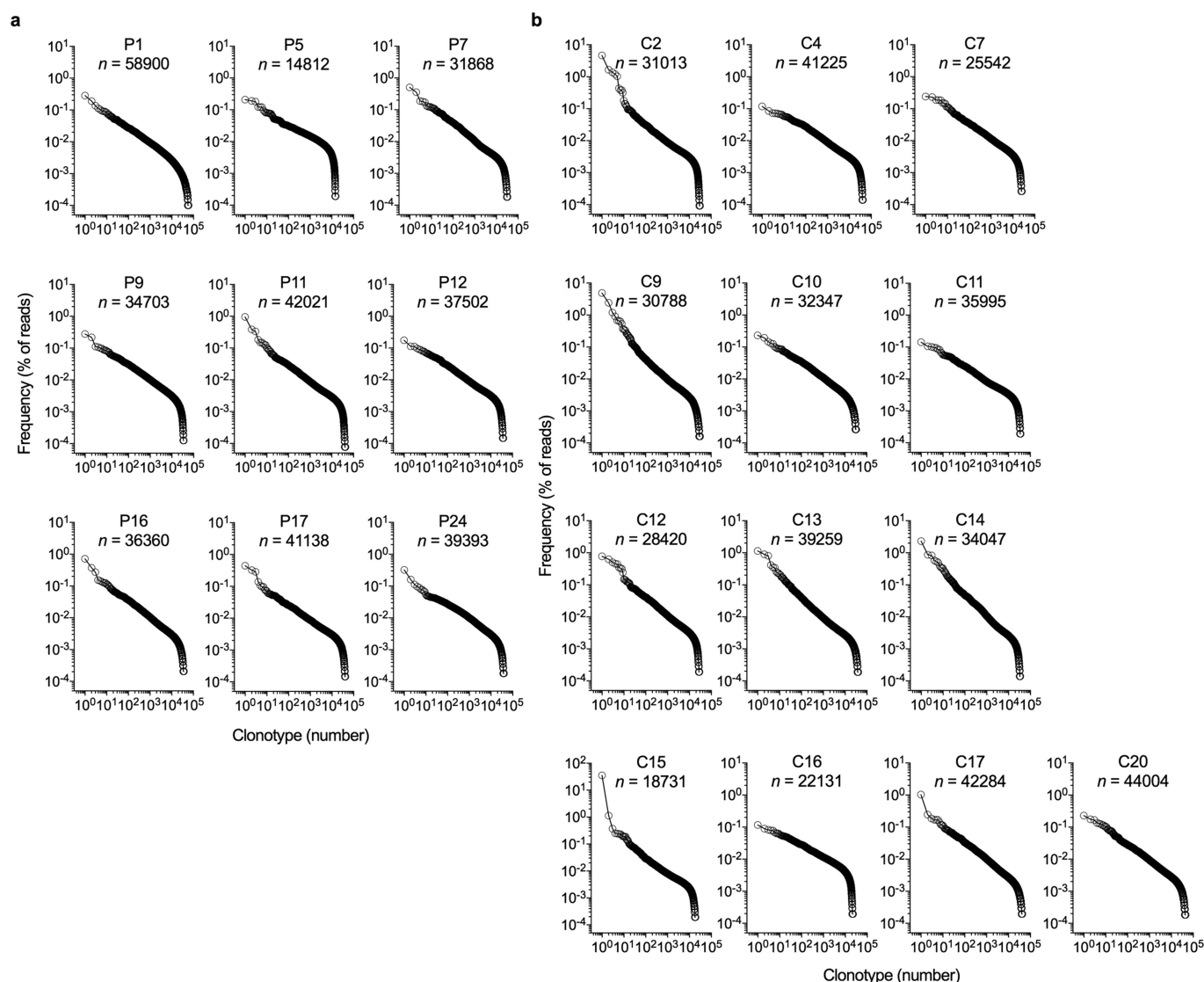
Extended Data Fig. 3 | Cytokine production and gene expression by autoreactive CD4⁺ T cell clones from patients with narcolepsy.

a, b, HCRT-specific (**a**, $n = 26$ biologically independent samples) and TRIB2-specific (**b**, $n = 18$ biologically independent samples) CD4⁺ T cell clones derived from patients with narcolepsy were stimulated with HCRT peptide pool (**a**) or TRIB2 peptide pool (**b**) presented by irradiated autologous B cells. Cytokines released in the 48-h culture

supernatants were quantified by bead-based multiplex assay. Data represent the mean \pm s.e.m. **c, d**, mRNA expression levels for 579 genes in HCRT-specific (**c**, $n = 6$ biologically independent samples; $n = 3$ from P1 and $n = 3$ from P8) and TRIB2-specific (**d**, $n = 6$ biologically independent samples; $n = 4$ from P22 and $n = 2$ from P14) CD4⁺ T cell clones were measured using NanoString technology. The top 50 expressed genes are shown. Data represent the mean \pm s.e.m.

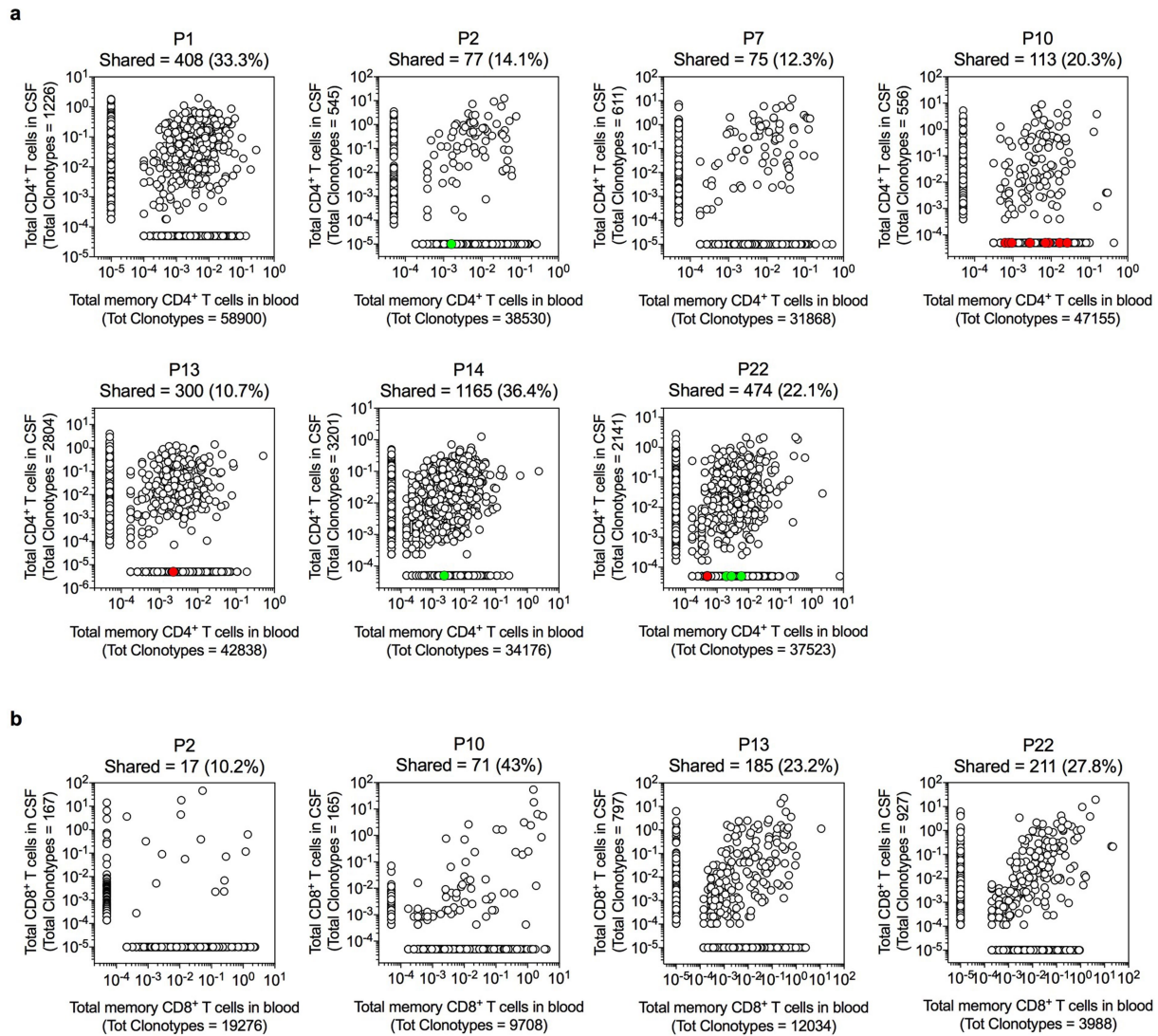


Extended Data Fig. 4 | Evaluation of T cell cross-reactivity with influenza virus antigens and MHC restriction of autoreactive T cell clones. **a**, **b**, HCRT-specific (**a**, $n = 61$ biologically independent samples) and TRIB2-specific (**b**, $n = 14$ biologically independent samples) CD4⁺ T cell clones from patients with narcolepsy were stimulated in the presence of irradiated autologous B cells pulsed with HCRT peptide pool, TRIB2 peptide pool or influenza vaccine (Influvac) and—for some clones ($n = 29$, HCRT-specific and $n = 9$, TRIB2-specific)—with haemagglutinin (HA) peptide pool. Each clone represents an individual clonotype. The c.p.m. values indicate the proliferation of autoreactive T cell clones measured after a 16-h pulse with [³H]thymidine are shown. **c**, The proliferative response of HCRT-specific, HLA-DR-restricted CD4⁺ T cell clones ($n = 9$ biologically independent samples) after stimulation with irradiated autologous B cells or a DRB1*15:01-transfected B cell line in absence or presence of HCRT peptide pool, is shown. Proliferation was measured on day 3 after a 16-h pulse with [³H]thymidine.



Extended Data Fig. 5 | Clonotypic analysis of blood memory CD4⁺ T cells. a, b, TCR V β sequencing was performed on memory CD4⁺ T cells ex vivo after sorting from peripheral blood of the indicated patients with narcolepsy (a) and healthy controls (b). The frequency distribution of all TCR V β clonotypes (n indicates total number of clonotypes) is shown.

In these samples, no sequences of autoreactive T cell clones were found. The number of sequenced clonotypes was comparable in patients and controls; a range of 14,812–58,900 ($39,494 \pm 10,514$ (mean \pm s.d.)) and a range of 18,731–44,004 ($33,354 \pm 7,831$ (mean \pm s.d.)), respectively. TCR sequencing data are available as Supplementary Information.



Extended Data Table 1 | Patients and control donors included in this study

Narcolepsy type 1 patients (NT1)									
ID	Gender	Age	Disease duration*	Cataplexy	HLA-DQB1*06:02	HLA-DRB1*15:01	HLA-DRB5*01:01	Treatment	CSF HCRT (pg/mL)
P1	M	21	3	Yes	+	+	+	Modafinil	64
P2	M	28	12	Yes	+	+	+	Modafinil	0
P3	F	16	4	Yes	+	+	+	Modafinil, Fluoxetine	31
P5	F	18	6	Yes	+	+	+	Methylphenidate	N.A.
P6	M	24	3	Yes	+	+	+	Modafinil, Fluoxetine	<20
P7	M	28	7	Yes	+	+	+	Modafinil	25
P8	M	48	7	Mild	+	+	+	Methylphenidate, Venlafaxine, Agomelatine	N.A.
P10†	F	44	27	Yes	+	+	+	Sodium Oxybate, Venlafaxine, Natalizumab	0
P11	M	32	14	Yes	—	—	—	Venlafaxine	0
P12	M	29	11	Yes	+	+	+	Methylphenidate, Venlafaxine	0
P13	F	21	2	Yes	+	+	+	Modafinil	96
P16	M	53	9	Yes	+	+	+	Modafinil, Bupropion	N.A.
P17	F	52	36	Yes	+	+	+	Ephredine	0
P20	F	46	12	Yes	+	+	+	Modafinil, Fluoxetine	<20
P22	M	25	2	Yes	—	—	—	None	<20
P24	M	35	1	Yes	+	+	+	Modafinil	68
Narcolepsy type 2 patients (NT2)									
P4‡	M	38	5	Mild	—	—	—	Methylphenidate, Venlafaxine	128
P9	F	51	31	No	—	—	—	Modafinil	N.A.
P14	F	43	2	No	—	—	—	Modafinil	395
Control donors									
C2	F	51	—	—	+	+	+	—	—
C4§	F	45	—	—	+	+	+	—	—
C7	N.A.	N.A.	—	—	+	+	+	—	—
C9	N.A.	N.A.	—	—	+	+	+	—	—
C10	N.A.	N.A.	—	—	+	+	+	—	—
C11	N.A.	N.A.	—	—	+	+	+	—	—
C12	N.A.	N.A.	—	—	+	+	+	—	—
C13	N.A.	N.A.	—	—	+	+	+	—	—
C14	N.A.	N.A.	—	—	+	—	—	—	—
C15	N.A.	N.A.	—	—	+	+	+	—	—
C16	N.A.	N.A.	—	—	+	+	+	—	—
C17	N.A.	N.A.	—	—	+	+	+	—	—
C20	N.A.	N.A.	—	—	+	+	+	—	—

N.A., not available.

*Duration in years.

†Co-morbidity, multiple sclerosis.

‡At a subsequent visit, nocturnal sleep of patient P4 deteriorated and cataplexy episodes were more frequent.

§First-degree family member of a patient with NT1.

Extended Data Table 2 | Epitope mapping and HLA restriction of autoreactive CD4⁺ T cell clones from patients with narcolepsy

Clonotype	PT	Specificity	Epitope (aa)	Restriction
1	P1	HCRT	17-36	HLA-DR
2	P1	HCRT	113-127	HLA-DR
3	P1	HCRT	17-36	HLA-DR
4	P1	HCRT	53-67 / 85-99	HLA-DR
5	P1	HCRT	17-31	HLA-DR
6	P1	HCRT	73-91	HLA-DR
7	P1	HCRT	N.D.	HLA-DR
8	P1	HCRT	49-78	HLA-DQ
9	P1	HCRT	17-36	HLA-DR
10	P1	HCRT	17-31	HLA-DR
11	P1	HCRT	105-119	N.D.
12	P1	HCRT	81-95	HLA-DQ
13	P3	HCRT	105-124	HLA-DR
14	P3	HCRT	73-87	HLA-DR
15	P3	HCRT	105-127	HLA-DP
16	P3	HCRT	N.D.	HLA-DR
17	P4	HCRT	57-76	N.D.
18	P4	HCRT	97-116	HLA-DR
19	P4	HCRT	73-92	HLA-DR
20	P4	HCRT	57-84	HLA-DR
21	P4	HCRT	97-116	HLA-DR
22	P4	HCRT	57-84	HLA-DR
23	P4	HCRT	65-84	HLA-DR
24	P8	HCRT	57-75	HLA-DR
25	P8	HCRT	57-76	HLA-DR
26	P8	HCRT	57-71	HLA-DR
27	P8	HCRT	57-71	HLA-DR
28	P8	HCRT	N.D.	N.D.
29	P8	HCRT	57-71	HLA-DR
30	P8	HCRT	29-43	HLA-DQ
31	P10	HCRT	69-87	HLA-DR
32	P10	HCRT	47-68	HLA-DR
33	P10	HCRT	45-59	HLA-DR
34	P10	HCRT	69-87	HLA-DR
35	P10	HCRT	33-52	HLA-DR
36	P10	HCRT	N.D.	HLA-DR
37	P10	HCRT	87-108	HLA-DR
38	P10	HCRT	25-44	HLA-DR
39	P10	HCRT	97-116	HLA-DR
40	P10	HCRT	69-91	HLA-DR
41	P10	HCRT	9-28	HLA-DR
42	P10	HCRT	65-84	HLA-DR
43	P10	HCRT	65-84	HLA-DR
44	P10	HCRT	69-91	HLA-DR
45	P10	HCRT	69-87	HLA-DR
46	P10	HCRT	69-87	HLA-DR
47	P10	HCRT	33-47/69-95/113-127	HLA-DR
48	P10	HCRT	111-131	HLA-DR
49	P10	HCRT	17-36	HLA-DR
50	P10	HCRT	97-116	HLA-DR
51	P10	HCRT	65-92	HLA-DR
52	P10	HCRT	69-97	N.D.
53	P10	HCRT	73-87	HLA-DR
54	P10	HCRT	69-91	HLA-DR
55	P10	HCRT	73-87	HLA-DR
56	P10	HCRT	69-91	HLA-DR
57	P10	HCRT	69-87	HLA-DR
58	P10	HCRT	73-87	HLA-DR
59	P10	HCRT	N.D.	N.D.
60	P11	HCRT	81-95	HLA-DR
61	P13	HCRT	25-44	HLA-DR
62	P14	HCRT	N.D.	HLA-DR
63	P14	HCRT	N.D.	HLA-DR
64	P22	HCRT	17-36	HLA-DR
65	P7	TRIB2	1-20	HLA-DP
66	P7	TRIB2	1-20	HLA-DP
67	P14	TRIB2	257-284	HLA-DQ
68	P14	TRIB2	177-196	HLA-DQ
69	P14	TRIB2	257-284	HLA-DQ
70	P14	TRIB2	257-284	HLA-DQ
71	P14	TRIB2	257-284	HLA-DQ
72	P14	TRIB2	1-20	HLA-DQ
73	P22	TRIB2	153-172	HLA-DR
74	P22	TRIB2	113-132	HLA-DR
75	P22	TRIB2	185-212	HLA-DQ
76	P22	TRIB2	265-284	HLA-DR
77	P22	TRIB2	49-68	HLA-DR
78	P22	TRIB2	313-332	HLA-DQ
79	P22	TRIB2	265-284	HLA-DR

N.D., not determined.

Extended Data Table 3 | HLA typing of patients with narcolepsy included in this study

ID	HLA- DRB1*	HLA- DRB1*	HLA- DQB1*	HLA- DQB1*	HLA- DQA1*	HLA- DQA1*	HLA- DPB1*	HLA- DPB1*	HLA- A*	HLA- A*	HLA- B*	HLA- B*	HLA- C*	HLA- C*
P1	01:01	15:01	05:01	06:02	01:01	01:02	04:01	13:01	03:01	11:01	07:02	35:01	04:01	07:02
P2	15:01	16:01	05:02	06:02	01:02	01:02	04:01	13:01	01:01	11:01	07:02	55:01	03:03	07:02
P3	11:01	15:01	03:01	06:02	01:02	05:05	04:01	20:01	32:01	68:01	27:07	44:02	07:04	15:02
P5	04:01	15:01	03:01	06:02	01:02	03:03	04:01	04:01	03:01	24:02	07:02	39:01	07:02	12:03
P6	15:01	16:01	05:02	06:02	01:02	01:02	10:01	11:01	01:01	02:01	07:02	55:01	03:03	07:02
P7	15:01	16:01	05:02	06:02	01:02	01:02	03:01	04:01	03:01	31:01	07:02	51:01	07:02	15:02
P8	11:04	15:01	03:01	06:02	01:02	05:05	02:01	04:01	25:01	25:01	18:01	44:05	02:02	12:03
P10	11:04	15:01	03:01	06:02	01:02	05:05	04:01	04:02	24:02	29:01	44:03	51:01	12:03	16:01
P11	01:01	04:07	03:01	05:01	01:01	03:03	02:01	04:01	02:01	11:01	07:02	35:01	04:01	07:02
P12	08:01	15:01	04:02	06:02	01:02	04:01	03:01	04:01	01:01	02:01	07:02	51:01	07:02	14:02
P13	13:02	15:01	06:02	06:04	01:02	01:02	02:01	02:01	01:01	24:02	39:06	58:01	07:02	07:01
P16	13:02	15:01	06:02	06:04	01:02	01:02	04:01	15:01	03:01	24:02	07:02	15:17	07:01	07:02
P17	04:07	15:01	03:01	06:02	03:03	01:02	04:01	04:01	11:01	11:01	35:01	56:01	01:02	04:01
P20	04:07	15:01	03:01	06:02	01:02	03:03	04:01	04:01	02:01	03:01	07:02	44:02	07:02	07:04
P22	01:01	16:01	05:01	05:02	01:01	01:02	04:01	14:01	03:01	24:02	35:01	37:01	04:01	06:02
P24	11:04	15:01	03:01	06:02	05:05	01:02	03:01	04:01	01:01	23:01	07:02	49:01	07:02	07:01
P4	04:01	13:01	03:01	06:03	01:03	03:03	04:01	04:01	11:01	30:02	15:01	44:02	03:04	05:01
P9	03:01	08:01	02:01	04:02	04:01	05:01	02:01	04:01	01:01	03:01	07:02	55:01	03:03	07:02
P14	01:02	04:02	03:02	05:01	01:01	03:01	02:01	04:01	02:01	29:01	44:03	45:01	16:01	16:01

Extended Data Table 4 | Summary of memory CD4⁺ T cell screenings

ID	Ex vivo	T cell library	HCRT-T cell reactivity	Isolation of HCRT-reactive clones	TRIB2-T cell reactivity	Isolation of TRIB2-reactive clones
P1	√	√	Yes	Yes	N.D.*	N.D.
P2	N.D.	√	Yes	None	Yes	None
P3	√	√	Yes	Yes	N.D.	N.D.
P5	√	√	Yes	None	Yes	None
P6	N.D.	√	Yes	None	N.D.	N.D.
P7	N.D.	√	Yes	None	Yes	Yes
P8	√	N.D.	Yes	Yes	N.D.	N.D.
P10	√	√	Yes	Yes	Yes	None
P11	√	√	Yes	Yes	None	N.D.
P12	N.D.	√	Yes	None	None	N.D.
P13	√	√	Yes	Yes	None	N.D.
P16	N.D.	√	Yes	None	Yes	None
P17	N.D.	√	Yes	None	Yes	None
P20	N.D.	√	Yes	None	None	N.D.
P22	√	√	Yes	Yes	Yes	Yes
P24	√	√	Yes	None	None	N.D.
P4	N.D.	√	Yes	Yes	N.D.	N.D.
P9	N.D.	√	Yes	None	N.D.	N.D.
P14	N.D.	√	Yes	Yes	Yes	Yes

*N.D., not done.

Necroptosis microenvironment directs lineage commitment in liver cancer

Marco Seehawer^{1,2,14}, Florian Heinzmann^{1,2,14}, Luana D'Artista^{1,2}, Jule Harbig^{1,2}, Pierre-François Roux^{3,4,5}, Lisa Hoenicke^{1,2}, Hien Dang⁶, Sabrina Klotz^{1,2}, Lucas Robinson^{3,4,5}, Grégory Doré^{3,4,5}, Nir Rozenblum³, Tae-Won Kang^{1,2}, Rishabh Chawla^{1,2}, Thorsten Buch⁷, Mihael Vucur⁸, Mareike Roth⁹, Johannes Zuber⁹, Tom Luedde⁸, Bence Sipos¹⁰, Thomas Longerich¹¹, Mathias Heikenwälder¹², Xin Wei Wang⁶, Oliver Bischof^{3,4,5} & Lars Zender^{1,2,13*}

Primary liver cancer represents a major health problem. It comprises hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC), which differ markedly with regards to their morphology, metastatic potential and responses to therapy. However, the regulatory molecules and tissue context that commit transformed hepatic cells towards HCC or ICC are largely unknown. Here we show that the hepatic microenvironment epigenetically shapes lineage commitment in mosaic mouse models of liver tumorigenesis. Whereas a necroptosis-associated hepatic cytokine microenvironment determines ICC outgrowth from oncogenically transformed hepatocytes, hepatocytes containing identical oncogenic drivers give rise to HCC if they are surrounded by apoptotic hepatocytes. Epigenome and transcriptome profiling of mouse HCC and ICC singled out *Tbx3* and *Prdm5* as major microenvironment-dependent and epigenetically regulated lineage-commitment factors, a function that is conserved in humans. Together, our results provide insight into lineage commitment in liver tumorigenesis, and explain molecularly why common liver-damaging risk factors can lead to either HCC or ICC.

Chronic liver inflammation and liver cirrhosis represent the most important risk factors for the development of primary liver cancer. Primary liver cancer comprises hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC), which are distinct with respect to their morphology, metastatic capacity and their response to cancer therapy^{1,2}. Whereas HCC grow in a solid, trabecular and sometimes pseudoglandular pattern with a high density of tumour cells, ICC are composed of ductular, papillary or solid tumour structures embedded in a dense tumour stroma, a feature shared with ontogenetically related pancreatic ductal adenocarcinomas³. HCC mostly show a local invasive growth restricted to the liver, whereas ICC tend to metastasize early, and also to distant organs⁴. Liver cells contain a high grade of cellular plasticity and it has been suggested that HCC and ICC both can derive from hepatocytes^{5,6}.

Epidemiological data indicate that in western countries, common liver-damaging risk factors such as obesity, alcohol abuse, dyslipidaemia, metabolic syndrome and steatohepatitis predispose to HCC as well as ICC development^{7,8}. However, so far the molecular mechanisms that determine HCC growth in some patients and ICC growth in others remain elusive.

Here we report that the hepatic microenvironment determines lineage commitment in liver tumorigenesis via epigenetic regulation. Using mosaic mouse models, we demonstrate that a necroptosis-associated hepatic cytokine microenvironment switches HCC to ICC development, independently of the oncogenic drivers. Pharmacological or genetic suppression of necroptosis revert the necroptosis-dependent cytokine microenvironment and switches ICC to HCC. Epigenome and

transcriptome profiling of mouse HCC and ICC identified *Tbx3* and *Prdm5* as major microenvironment-dependent and epigenetically regulated lineage-commitment factors, which were validated in a cohort of 199 cases of human HCC and ICC. Our study provides fundamental insights into how lineage commitment in liver tumorigenesis is regulated, and explains how common hepatic risk factors such as a western lifestyle and fatty liver disease can lead to the development of either HCC or ICC.

Generation of HCC or ICC using transposon mouse models

To study liver tumorigenesis in mice, we used a well-established mouse model in which transposable elements are stably delivered into the liver via hydrodynamic tail-vein (HDTV) injection^{9,10}. To model the frequent upregulation of MYC and the induction of the MEK-ERK or PI3K-mTOR signalling in human HCC¹¹, we engineered transposon vectors co-expressing oncogenic mouse *Myc* and human *NRAS*^{G12V} (pCaMIN) or mouse *Myc* and *Akt1* (pCaMIA) (Extended Data Fig. 1a). When HDTV was used to co-deliver these vectors together with a sleeping beauty transposase (SB13)-encoding plasmid (pSB13) into the hepatocytes of *p19^{Arf}-/-* (encoded by the *Cdkn2a* locus) mice, we observed multifocal liver carcinomas (Fig. 1a), which resembled HCC with a solid and sometimes trabecular growth pattern and steatosis (Extended Data Fig. 1b). Immunohistochemistry revealed strong expression of the hepatocyte-specific transcription factor hepatocyte nuclear factor-4α (HNF4α) (Fig. 1b) but a lack of biliary type keratin 19 (K19) expression (Fig. 1c), therefore confirming all pCaMIN- and pCaMIA-induced tumours to be HCC (Extended Data Fig. 1d).

¹Department of Internal Medicine VIII, University Hospital Tuebingen, Tuebingen, Germany. ²Department of Physiology I, Institute of Physiology, Eberhard Karls University Tuebingen, Tuebingen, Germany. ³Institut Pasteur, Nuclear Organization and Oncogenesis Unit, Department of Cell Biology and Infection, Paris, France. ⁴INSERM, U993, Paris, France. ⁵Equipe Labellisée Fondation ARC pour la recherche sur le cancer, Villejuif, France. ⁶Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. ⁷Institute of Laboratory Animal Science University of Zurich, University of Zurich, Schlieren, Switzerland. ⁸RWTH University Hospital Aachen, Department of Gastroenterology, Digestive Diseases and Intensive Care Medicine (Department of Medicine III), Aachen, Germany. ⁹Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria. ¹⁰Institute of Pathology, University of Tuebingen, Tuebingen, Germany. ¹¹Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany. ¹²Division of Chronic Inflammation and Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹³Translational Gastrointestinal Oncology Group, German Consortium for Translational Cancer Research (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁴These authors contributed equally: Marco Seehawer, Florian Heinzmann. *e-mail: lars.zender@med.uni-tuebingen.de

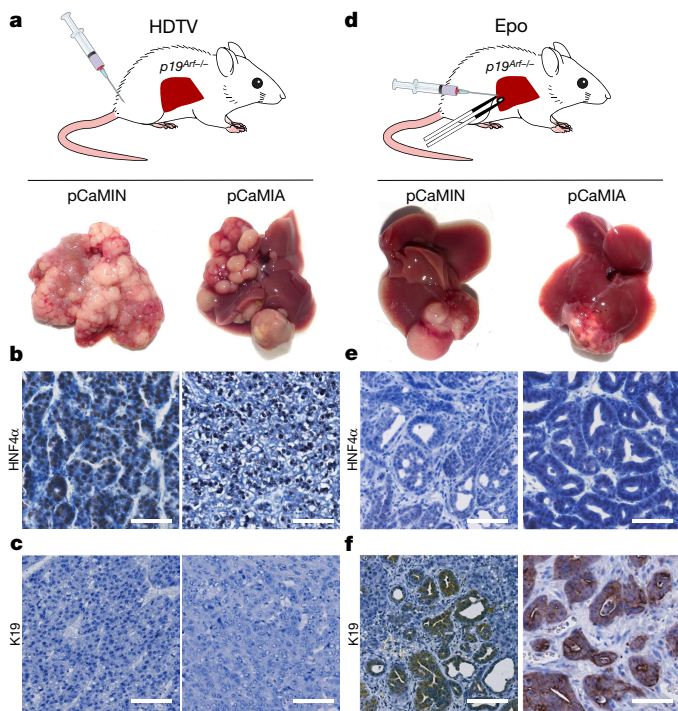


Fig. 1 | Intrahepatic delivery of transposable elements encoding *Myc* and *NRAS*^{G12V} or *Myc* and *Akt1* into *p19Arf*^{-/-} mice results in multifocal HCC or unilocal ICC. **a**, Intrahepatic delivery of the transposable vectors pCaMIN (encoding *Myc* and *NRAS*^{G12V}; compare Extended Data Fig. 1a) ($n = 11$) or pCaMIA (encoding *Myc* and *Akt1*; compare Extended Data Fig. 1b) ($n = 14$) via HDTV results in multifocal tumour development after 4 weeks. **b**, **c**, Representative micrographs of immunohistochemistry staining against K19 (**b**) and HNF4 α (**c**). Scale bar, 100 μ m. **d**, Epo treatment of pCaMIN (compare Extended Data Fig. 1a) ($n = 19$) or pCaMIA (compare Extended Data Fig. 1b) ($n = 8$) results in the development of unilocal liver carcinomas 4 weeks after electroporation. **e**, **f**, Representative immunohistochemistry images of K19 (**e**) and HNF4 α (**f**). Scale bars, 100 μ m. Experiments were conducted in three independent cohorts.

Development of multifocal HCC throughout the liver is a situation often seen in patients with advanced-stage HCC. However, to study unilocal hepatic hepatocellular carcinomas, we first injected pCaMIN or pCaMIA under the liver capsule and then applied *in vivo* electroporation (Epo), which efficiently and focally transfects hepatocytes¹² (Fig. 1d). Histopathological and immunohistochemistry analysis revealed that most tumours were either ICC or combined ICC–HCC (Extended Data Fig. 1c, e) developing focally at the electroporation site (Fig. 1g), staining negatively or weakly and focally restricted for HNF4 α (Fig. 1e) but strongly positive for K19 (Fig. 1f).

In vivo lineage tracing reveals hepatic origin of ICC

Recent studies suggested that ICC can be derived from cholangiocytes, liver progenitor cells or hepatocytes^{5,13}. To determine the cell of origin of tumorigenesis induced by Epo, we used a well-established lineage-tracing mouse model, in which only differentiated hepatocytes show a switch from red to green fluorescence¹⁴. These mice were crossed on a *p19Arf*-deficient background to obtain *ROSA*^{mt/mG} \times *Alb-cre* \times *p19Arf*^{-/-} mice (Fig. 2a, Extended Data Fig. 1f). Next, we stably delivered pCaMIN transposon vectors by either HDTV or Epo. Whereas HDTV of pCaMIN again triggered the growth of HCC (HNF4 α ⁺, K19⁻), Epo delivery still resulted in the development of ICC (HNF4 α ⁻, K19⁺) (Fig. 2b, Extended Data Fig. 1g, h). Native fluorescence analysis of HDTV-induced HCC and Epo-triggered ICC revealed homogenous green fluorescence throughout the tumours, thereby identifying differentiated hepatocytes as the cells of origin (Fig. 2c) for both tumour types. Co-localization of native GFP (green)

and K19 (red; Extended Data Fig. 1i) confirmed the hepatocytic origin of K19-positive ICC. The ICC typical tumour stroma only showed red fluorescence (Fig. 2c).

Microenvironment determines lineage commitment

We next sought to address why HDTV-mediated transposon delivery leads to HCC, but delivery of the same vectors via Epo leads to ICC. To exclude the role of quantitative differences in transposon integration, we injected pCaMIN via HDTV followed by a mock Epo treatment (without DNA) of a defined area of the same liver two hours later (Fig. 3a). Histopathological analyses identified developing tumours as ICC or combined ICC–HCC (Fig. 3b, c), whereas HDTV-induced tumours were pure HCC (Fig. 3b, d). Quantitative PCR (qPCR) analyses with transposon-specific primers on HDTV-mediated HCC and Epo-mediated ICC ruled out the possibility that different levels of transposon integration contribute to the distinct tumour phenotypes (Extended Data Fig. 1j, k).

We hypothesized that liver electroporation might cause genetic mutations that drive cholangiocytic lineage determination. Indeed, several genes and pathways have been reported to affect lineage commitment in liver tumorigenesis¹³. Therefore, we conducted laser capture microdissection and ‘purified’ HCC versus ICC tissues from mice containing pCaMIN Epo-induced mixed ICC–HCC tumours (Extended Data Fig. 2a). Apart from the engineered *NRAS*^{G12V} mutation we found 12 recurrent mutations (Extended Data Fig. 2b) including mutations in the hydrolase function gene *Car7* or the glycoprotein *Dag1*; however, both mutations were found in HCC and ICC samples. Notably, *Fam72a* mutations were exclusively found in two HCC samples, and *Smc3* mutations were exclusively found in two ICC samples. Yet, interrogation of the COSMIC and cBioPortal databases did not reveal specificity of mutations for HCC and ICC, respectively, and functional testing of mutated *Fam72a* in our mouse model had no effect on lineage commitment (Extended Data Fig. 2c, d).

Necroptosis-dependent microenvironment leads to ICC

Next, we investigated whether HDTV and Epo treatment differentially shape specific hepatic microenvironments to determine HCC or ICC. We analysed pCaMIN-injected livers with HDTV or Epo within the first 5 days, a time in which lineage commitment can be distinguished in precursor lesions (HNF4 α ⁻, K19⁺ microcarcinomas after Epo and HNF4 α ⁺, K19⁻ microcarcinomas after HDTV; Extended Data Fig. 3a).

Three days after HDTV or Epo treatment, both methods induce eosinophilic areas of damaged tissue with an associated inflammatory reaction (Extended Data Fig. 3b). Hepatic stellate cells are important cells that shape the hepatic microenvironment after liver damage¹⁵. However, immunostaining analysis of α SMA did not reveal differences between HDTV- and Epo-treated livers (Extended Data Fig. 4a). Similarly, the quantification of Kupffer cells did not reveal differences between the two groups (Extended Data Fig. 4b), and clodronate-mediated depletion of Kupffer cells did not affect tumour phenotype (Extended Data Fig. 4c).

Next, we performed immunohistochemistry for T cells (CD3), monocytes and (neutrophilic) granulocytes (Ly6G) as well as B cells (B220) and antigen-presenting cells (MHCII) to compare inflammatory infiltrates in HDTV- versus Epo-treated livers (Fig. 4a, b). Notably, no significant differences between HDTV- and Epo-treated mouse livers could be observed (Extended Data Fig. 4d). Further quantifications of immune cells via flow cytometry did not reveal changes in the numbers of T cells (helper or killer) or immature myeloid cells (monocytic or neutrophilic) or macrophages (Extended Data Fig. 4e, f).

We then applied TUNEL staining, which detects several types of cell death including necrosis, apoptosis and necroptosis¹⁶. TUNEL staining did not reveal any differences in the overall number of dying cells in mouse livers transfected with HDTV (69% \pm 16%; mean \pm s.d.) or Epo (70% \pm 18%) (Fig. 4c). Dying cells were found to be hepatocytes as the TUNEL signal (red fluorescence) co-localized

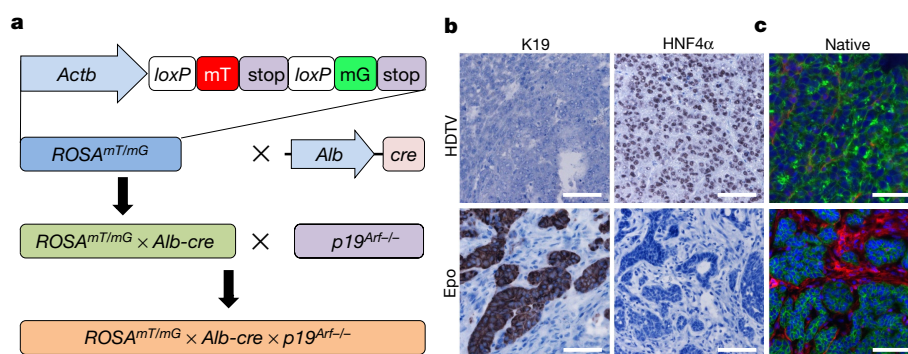


Fig. 2 | In vivo lineage tracing identifies hepatocytes as cells of origin for ICC development. **a**, $ROSA^{mT/mG}$ mice were crossed to $Alb-cre$ mice. The resulting mice were intercrossed with $p19^{Arf-/-}$ mice to generate $ROSA^{mT/mG} \times Alb-cre \times p19^{Arf-/-}$ mice. **b**, **c**, Representative images

of tumours 4 weeks after HDTV (top) or Epo (bottom) treatment of the pCaMIN vector in $ROSA^{mT/mG} \times Alb-cre \times p19^{Arf-/-}$. Shown are immunohistochemistry staining results using antibodies against K19 and HNF4α (**b**) as well as native fluorescence (**c**) ($n = 4$). Scale bars, 100 μm.

with membrane-associated green fluorescent protein (GFP) of hepatocytes ($ROSA^{mT/mG} \times Alb-cre \times p19^{Arf-/-}$ mouse) (Extended Data Fig. 5a).

Apoptosis and necroptosis are two relevant forms of cell death in the pathogenesis of human liver disease¹⁷. To discriminate between these two types of cell death, we performed western blot analyses, which revealed high levels of the apoptosis marker cleaved caspase 3 in HDTV-treated livers but only low levels in Epo-treated livers, indicating that HDTV predominantly induces apoptosis (Fig. 4d, top). Necroptosis was described to be involved in the pathogenesis of different liver diseases such as non-alcoholic steatohepatitis or drug-induced liver injury^{17–19}. In these studies, necroptosis was mostly determined by quantification of RIPK3, a kinase that also has functions outside the necroptosis signalling cascade. Notably, Epo-treated livers showed high levels of phosphorylated MLKL and increased mRNA levels of *Ripk3* (Fig. 4d, bottom, Extended Data Fig. 5b), both biomarkers for necroptosis, when compared to HDTV-treated mouse livers. In addition,

immunohistochemistry staining of total RIPK3 and phosphorylated RIPK3 (pRIPK3) revealed positive signals in the electroporation area (Fig. 4e and Extended Data Fig. 5c).

Cells undergoing necroptosis release damage-associated molecular patterns that can shape the microenvironment via a pattern-recognition-receptor-dependent cytokine release in immune cells²⁰. Therefore, we conducted cytokine expression profiling revealing a strong induction of different cytokines in Epo- versus HDTV-treated livers (Fig. 4f). The most differentially regulated factors we identified included *Ccl4*, *Aimp1*, *Cxcl13*, *Ccl6*, *Ccl8*, *Pf4* and *Osm*.

To address whether these are causally linked to the high levels of necroptotic cell death in Epo-treated livers, we used necrostatin-1 (Nec-1), a very potent inhibitor of RIPK1²¹ and suppressor of necroptosis. Pre-treatment (starting 3 days before Epo) with Nec-1 for 3 days significantly reduced Epo-induced cell death as detected by TUNEL assay (Extended Data Fig. 6a, b), and shifted cell death towards apoptosis, as shown by caspase 3 cleavage (Extended Data Fig. 6c).

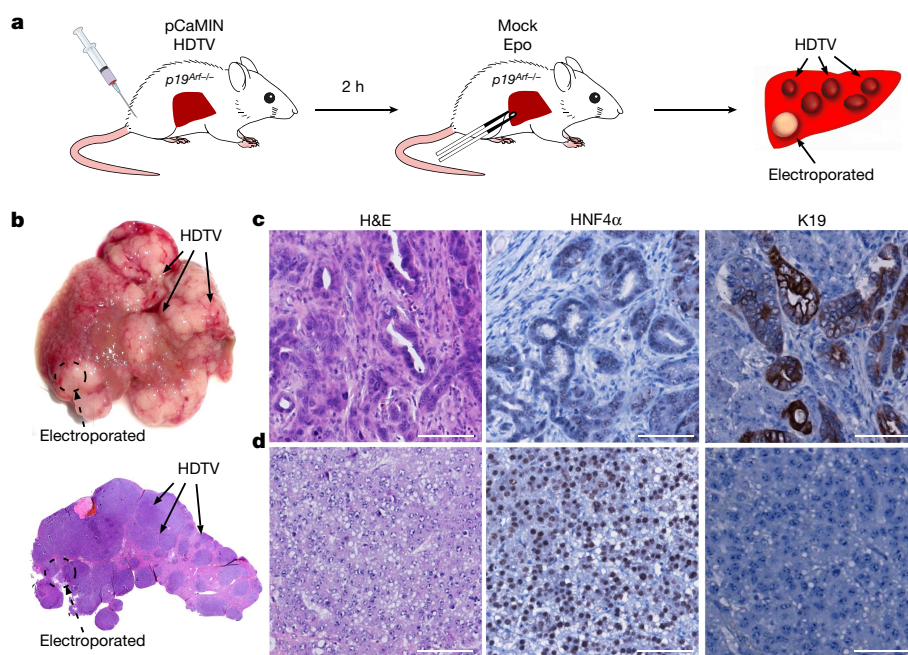


Fig. 3 | Electroporation associated microenvironment determines outgrowth of ICC from hepatocytes. **a**, Schematic experimental outline. Livers of $p19^{Arf-/-}$ mice were first hydrodynamically injected with pCaMIN and SB13 vectors and subsequently mock electroporated at a defined liver region. **b**, Macroscopic photograph of mouse liver (top) and corresponding representative haematoxylin and eosin (H&E) staining

(bottom) 3 weeks after HDTV and subsequent mock Epo treatment ($n = 3$). Scale bars, 500 μm. (**c**, **d**) Representative photographs of H&E- (left panel), HNF4α- (middle panel) and K19 stained (right panel) tumours in the mock electroporated area (scale bar, 100 μm) or (**d**) outside ($n = 3$) (scale bar, 100 μm).

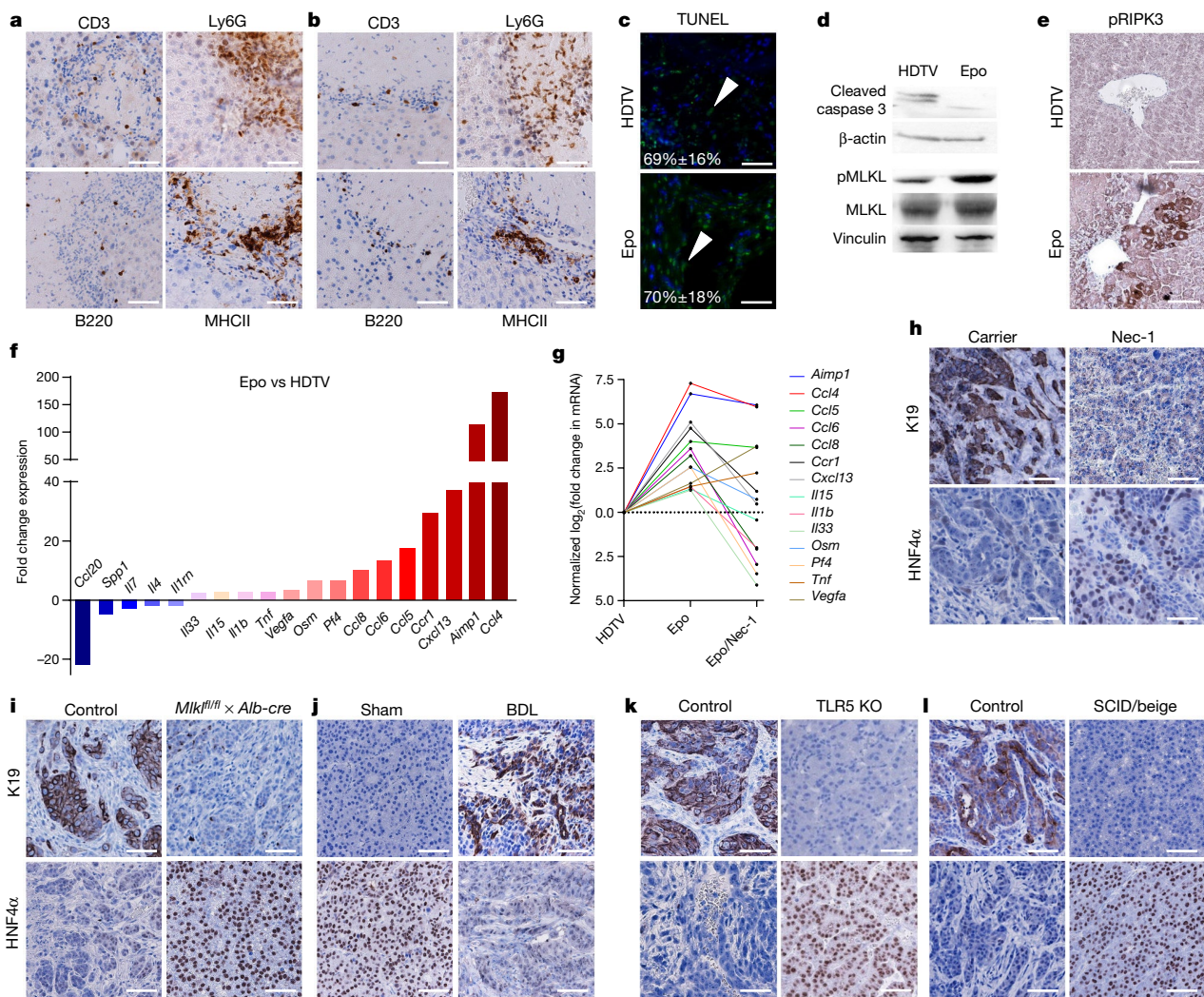


Fig. 4 | Necroptosis-dependent cytokine microenvironment determines cholangiocarcinoma development. **a, b**, Immunohistochemistry for CD3, Ly6G, B220 and MHCII 3 days after HDTV (**a**) or Epo (**b**) transfection of pCaMIN and SB13 into $p19^{Arf-/-}$ mice ($n = 4$). Scale bars, 100 μ m. **c**, TUNEL staining (green fluorescence, white arrowheads) and quantification on liver sections 3 days after pCaMIN transfection via HDTV (top) or Epo (bottom) ($n = 3$). Nuclei were counterstained blue with DAPI. Scale bars, 100 μ m. **d**, Western blot for apoptosis marker cleaved caspase 3 (top) and total or phosphorylated MLKL (pMLKL) (bottom) in liver lysates after transfection via HDTV or Epo ($n = 3$). **e**, Immunohistochemistry for pRIPK3 3 days after transfection via Epo or HDTV ($n = 3$ each). Scale bars, 100 μ m. **f**, Cytokine mRNA expression (fold change) in Epo- versus HDTV-treated livers after 3 days ($n = 2$). Data represent fold change of the mean from each group. **g**, Cytokine mRNA expression (fold induction) in Epo-treated versus HDTV-treated or Epo- and Nec-1-treated versus Epo-treated livers

Nec-1 treatment efficiently dampened necroptosis as shown by a reduced pMLKL signal (Extended Data Fig. 6d). Immune cell profiling revealed that Nec-1 treatment did not affect the numbers of intrahepatic B220-positive cells, CD3-positive cells and MHCII-positive cells (Extended Data Fig. 6e). A significant reduction was found only for Ly6G-positive cells in the Nec-1-treated versus carrier-treated mice after Epo treatment (Extended Data Fig. 6e). Notably, Nec-1-attenuated the induction of most Epo-associated cytokines (Fig. 4g) and switched ICC development towards the outgrowth of solid or trabecular HNF4 α -positive HCC tumours with no or only weak K19 expression (Fig. 4h and Extended Data Fig. 6f). By contrast, carrier-treated mice still revealed an ICC phenotype with K19-positive glandular structures and only microfocal HNF4 α -positive areas (Fig. 4h). To confirm the decisive role of necroptosis in liver

($n = 2$). Data are log₂ fold change of the mean from each group. **h**, Immunohistochemistry for K19 and HNF4 α of mice pre-treated with carrier (left) or Nec-1 (right) before pCaMIN Epo transfection ($n = 4$). Scale bars, 100 μ m. **i**, Immunohistochemistry for K19 and HNF4 α on liver tumour sections of $Mlkl^{fl/fl} \times Alb-cre^{-/-}$ (left) or $Mlkl^{fl/fl} \times Alb-cre^{+/+}$ (right) mice after pCaMIN Epo transfection ($n = 5$). Scale bars, 100 μ m. Compare to Extended Data Fig. 6m. **j**, Immunohistochemistry for K19 and HNF4 α . $p19^{Arf-/-}$ mice were subjected to HDTV and sham operation (control, left, $n = 3$), or bile duct ligation (BDL) and HDTV of pCaMIN (right, $n = 5$). Scale bars, 100 μ m. **k**, Immunohistochemistry for K19 and HNF4 α on liver sections from wild-type ($n = 4$) (left) or TLR knockout (KO; lacking TLR2, TLR3, TLR4, TLR7 and TLR9) ($n = 3$) mice (right) after pCaMIN Epo transfection. Scale bars, 100 μ m. **l**, Immunohistochemistry for K19 and HNF4 α on wild-type ($n = 4$) (left) or SCID/beige ($n = 8$) mice (right) after pCaMIN Epo transfection. Scale bars, 100 μ m.

cancer lineage commitment, we generated hepatocyte-specific *Mlkl* knockout mice by intercrossing *MLKL^{fl/fl}* mice²² with *Alb-cre* mice²³, resulting in *Alb-cre* \times *Mlkl^{fl/fl}* mice (MLKL Δ Hep). Western blot analysis of isolated hepatocytes confirmed hepatocyte-specific knockout of *Mlkl* (Extended Data Fig. 6g). Further, hepatic MLKL deficiency completely abolished the pMLKL signal 3 days after Epo treatment (Extended Data Fig. 6h).

Next, we subjected *Alb-cre^{+/-} × Mlkl^{fl/fl}* or *Alb-cre^{-/-} × Mlkl^{fl/fl}* mice to Epo treatment with pCaMIN together with a plasmid encoding Cas9n and a single guide RNA (sgRNA) against *p19^{Arf}*. CRISPR-induced p19 knockout robustly allowed for tumour development after pCaMIN delivery, and this was slightly delayed compared to the germline p19Arf knockout (Extended Data Fig. 6i). Hepatocyte-specific MLKL deficiency had no influence on the hepatic immune

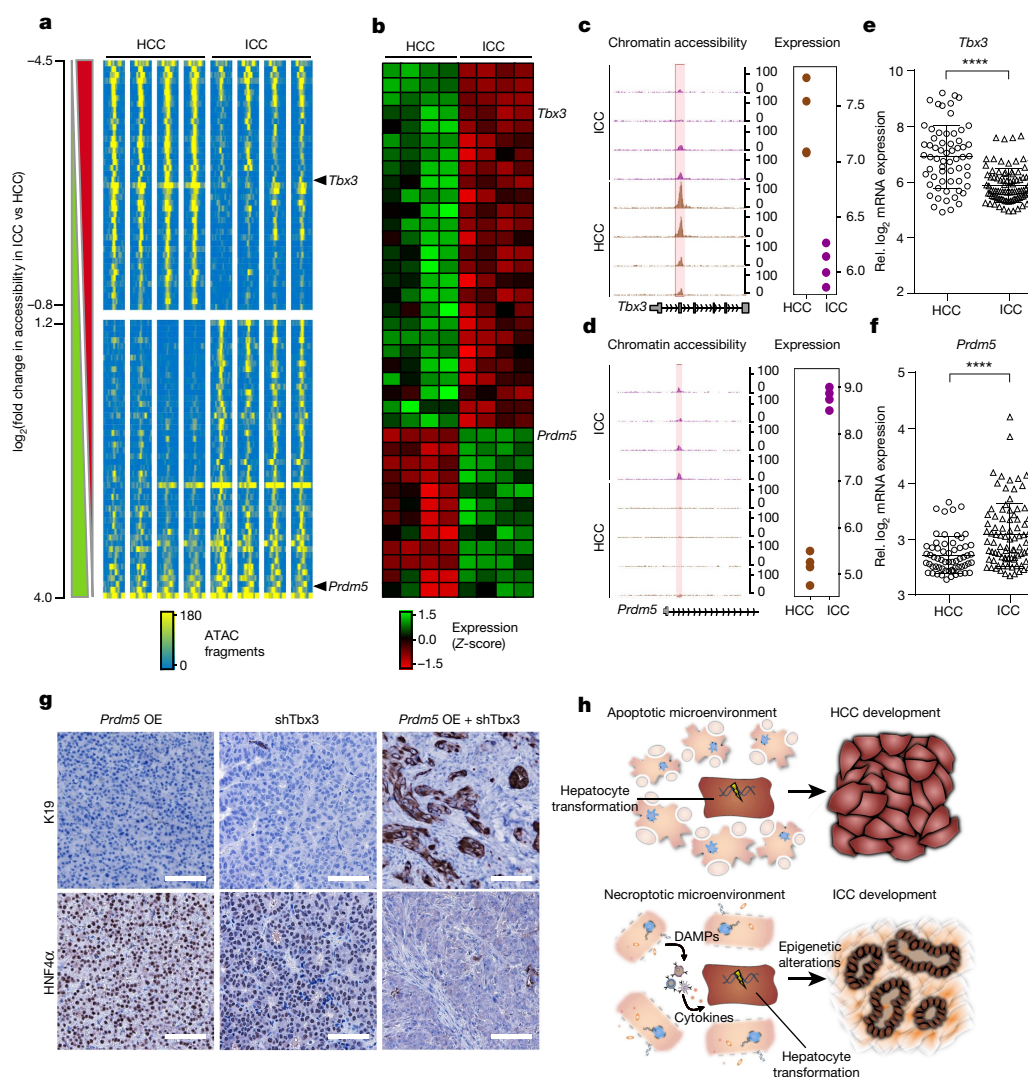


Fig. 5 | HCC and ICC derived from oncogenically transformed hepatocytes are defined by unique epigenetic signatures. a, ATAC-seq density heat map of chromatin regions that are differentially accessible between HCC and ICC. Peaks are ranked according to the fold-change in signal in normalized ATAC fragment counts in ICC versus HCC. The data are expressed as smoothed normalized fragment pseudocounts in 25-base-pair (bp) windows ± 1 kb around the centre of peaks. The lateral bars on the left depict whether the ATAC signal is significantly increased (green) or decreased (red) in ICC compared to HCC as assessed with EdgeR. *Tbx3*- and *Prdm5*-associated regulatory elements are indicated. For each transcription factor (TBX3 or PRDM5), $n = 4$ cases, two-sided moderated t -statistics. **b**, Heat map of transcriptome data comparing HCC and ICC showing differentially expressed probes. Probes matching *Tbx3* and *Prdm5* are indicated by arrows. For each transcription factor

(TBX3 or PRDM5), $n = 4$ cases, two-sided moderated t -statistics. **c**, **d**, Integrative analysis of chromatin accessibility (left) and transcriptome data (right) around *Tbx3* (**c**) and *Prdm5* (**d**) genes comparing HCC and ICC. Chromatin accessibility is expressed as smoothed, normalized fragment pseudo-counts in 100-bp windows. Absolute gene expression is represented in log scale. **e**, **f**, Gene expression in 199 human HCC and ICC of *TBX3* (**e**) and *PRDM5* (**f**). **** $P < 0.0001$, Student's two-sided t -test. Data are mean \pm s.d. **g**, Immunohistochemistry staining for K19 and HNF4 α of HDTV-derived tumours after co-delivery of pCaMIN and a *Prdm5*-overexpression transposon (*Prdm5* OE) ($n = 4$), pCaMIN plus *Tbx3* shRNA (shTbx3) ($n = 5$), or pCaMIN plus shTbx3 and *Prdm5* OE ($n = 3$) in *p19^{Arf}^{-/-}* mice. Scale bars, 100 μ m. **h**, Schematic representation of the proposed model. DAMPs, damage-associated molecular patterns.

cell infiltrates (Extended Data Fig. 6j); however, in line with our Nec-1 experiments, we found a reduction of Epo-associated cytokines (Extended Data Fig. 6k).

When MLKL Δ Hep mice were followed up for tumour development after pCaMIN Epo treatment, we found outgrowth of solid, HNF4 α -positive HCC, whereas control mice developed K19-positive ICC (*MLKL^{fl/fl} × Alb-cre^{-/-}* or *MLKL* wild-type \times *Alb-cre^{+/+}*; Fig. 4i and Extended Data Fig. 6l, m). Collectively, our data suggest that a necroptosis-enriched liver microenvironment promotes ICC outgrowth from oncogenically transformed hepatocytes.

Our findings were further validated in a mouse model of bile duct ligation-mediated liver damage, in which necroptotic cell death is prominent¹⁹ (Extended Data Fig. 6n, o). Notably, pCaMIN HDTV treatment after bile duct ligation induced the outgrowth of K19-positive and mostly HNF4 α -negative poorly differentiated liver carcinomas that

morphologically resemble tumours with mixed hepatobiliary features (Fig. 4j).

To address whether the identified mechanism of lineage determination also holds true in primary human liver tumorigenesis, we investigated mRNA expression of apoptosis ($n = 83$) or necroptosis ($n = 10$) related genes (Supplementary Table 4) in a cohort comprising 199 cases of HCC and ICC²⁴. Hierarchical clustering analysis revealed that a necroptosis signature is enriched in patients with ICC compared to those with HCC. By contrast, we found an 'apoptosis signature' in HCC (Extended Data Fig. 7a). Of note, *RIPK3* expression was significantly increased in ICC compared to HCC samples (Extended Data Fig. 7b).

To address the role of Toll-like receptors (TLRs), which can induce cytokine release after activation by necroptotic damage-associated molecular patterns, we subjected mice deficient for numerous TLRs²⁵ (TLR2, TLR3, TLR4, TLR7 and TLR9) to Epo-mediated pCaMIN

and Cas9n-*p19^{Arf}* sgRNA delivery. TLR deficiency did not prevent the induction of pMLKL 3 days after Epo treatment (Extended Data Fig. 7c). Similarly, TLR deficiency did not affect the numbers of liver-infiltrating immune cells (Extended Data Fig. 7d); however, TLR deficiency suppressed the induction of five of the six necroptosis-associated/MLKL-dependent (compare Extended Data Fig. 6k) cytokines or cytokine receptors (Extended Data Fig. 7e). Notably, TLR deficiency prevented ICC development but instead resulted in the outgrowth of HCC (Fig. 4k, Extended Data Fig. 7f). As the switch from ICC to HCC could also be observed in mice lacking only TLR2 and TLR4, our data suggest that these TLRs are the most crucial (Extended Data Fig. 7g).

We next electroporated pCaMIN together with Cas9n and sgRNA against *p19^{Arf}* into severe combined immunodeficient (SCID)/beige mice (impaired adaptive and innate immunity) or syngeneic control mice. Whereas tumours arising in control mice represented K19-positive but HNF4 α -negative ICC, immunocompromised SCID/beige mice developed HCC (Fig. 4l), suggesting that the crucial function of TLR signalling must be ascribed to TLR on immune cells.

Unique epigenetic signatures define HCC and ICC

Because we were unable to identify spontaneously acquired mutations that could account for ICC versus HCC outgrowth, we reasoned that lineage commitment might be epigenetically regulated. We established clonal cell lines from pCaMIN Epo-derived ICC and pCaMIN HDTV-derived HCC (Extended Data Fig. 8a). Cultured ICC and HCC gave rise to ICC or HCC after subcutaneous injection of tumour cells in immunodeficient *Rag2*^{-/-} mice, suggesting a stable maintenance of the tumour genotype during in vitro cultivation (Extended Data Fig. 8b, c).

We then profiled the chromatin accessibility landscape via assay for transposase-accessible chromatin using sequencing (ATAC-seq) in our cultured HCC and ICC cells. We generated fragment density heat maps ranked according to the signal fold change in ICC versus HCC to focus on differentially accessible chromatin regions (Fig. 5a). This allowed us to identify a total number of 108 chromatin regions that showed significant changes in chromatin accessibility, between HCC and ICC. *k*-means clustering analysis clearly separated ICC from HCC cell lines (Extended Data Fig. 8d). Next, we integrated our ATAC-seq data with transcriptome data that we generated from the same cell lines (Fig. 5b). These analyses pinpointed two genes, *Tbx3* and *Prdm5*, both transcription factors with described roles in carcinogenesis^{26,27}. Although a transcription start site (TSS)-proximal, intronic *Tbx3* chromatin region was mostly accessible in mouse HCC, it was inaccessible in ICC (Fig. 5c). The observed reciprocal patterns positively correlated with increased *Tbx3* mRNA expression in HCC when compared to ICC, a result we confirmed independently by quantitative PCR with reverse transcription (qRT-PCR) (Extended Data Fig. 8e).

As for *Prdm5*, we found the opposite regulation of chromatin accessibility when compared to *Tbx3* (Fig. 5d, Extended Data Fig. 8f). Notably, *Tbx3* and *Prdm5* gene expression levels in human HCC and ICC revealed the same patterns (Fig. 5e, f).

We hypothesized that lineage commitment in liver tumorigenesis is regulated by the expression of *Tbx3* and *Prdm5*. To consolidate this, we conducted functional genetic experiments. First, CaMIN transposon vectors also encoding a control short hairpin RNA (shRNA) were stably co-delivered via Epo with a transposon vector encoding full-length *Tbx3*. Outgrowing tumours revealed a partial shift from ICC to HCC, as they grew in a solid growth pattern of undifferentiated cells without glandular structures and completely lost tumour stroma and K19 positivity, but lacked HNF4 α expression (Extended Data Fig. 9a, left). Of note, these tumours are of hepatocytic origin as shown in our lineage-tracing model (Extended Data Fig. 9b). Notably, if in addition to *Tbx3* overexpression, *Prdm5* was suppressed by stable RNA interference (Extended Data Fig. 9c, shPrdm5_1), tumours gained differentiation and showed increased HNF4 α expression (Extended Data Fig. 9a, right). To corroborate the role of *Tbx3* and *Prdm5* further, we conducted the reverse experiment. We performed HDTV with

pCaMIN co-expressing a control shRNA (pCaMIN-shRen) or a *Tbx3* shRNA (pCaMIN-shTbx3) together with a transposon overexpressing full-length *Prdm5*.

Although, *Prdm5* overexpression or shRNA-mediated knockdown of *Tbx3* alone were not able to induce a switch from HCC to ICC, simultaneous *Prdm5* overexpression and *Tbx3* knockdown resulted in the outgrowth of HNF4 α -negative, K19-positive ICC (Fig. 5g). Together, these data strongly support the idea that TBX3 and PRDM5 act synergistically to determine lineage commitment in primary liver cancer.

To gain further mechanistic insights, we performed *Tbx3* and *Prdm5* chromatin-immunoprecipitation followed by sequencing (ChIP-seq) analysis in our HCC and ICC cell lines. There was a high concordance between ATAC-seq and *Tbx3* and *Prdm5* ChIP-seq signals in HCC and ICC (Extended Data Fig. 10a, b). The integration of ChIP-seq datasets with gene expression profiles obtained from HCC with stable *Tbx3* knockdown (Extended Data Figs. 9d, 10c) or ICC with stable *Prdm5* knockdown (Extended Data Figs. 9c, 10d) identified both direct transcriptional targets, such as *Tgfb2* or *Thbs1* for *Tbx3*, and *Col3a1* or *Ephb2* for *Prdm5*, and indirect targets, thus allowing us to delineate comprehensive downstream regulatory networks (Extended Data Fig. 10e–g, Supplementary Table 1).

Finally, to determine how *Tbx3* and *Prdm5* might be regulated, we conducted gene expression analysis of chromatin remodelling enzymes after Epo or HDTV treatment. Notably, in the Epo-treated livers we found upregulation of several chromatin remodellers (such as *Ciita*, *Hdac5* or *Ncoa1*) (Extended Data Fig. 10h), suggesting involvement in the epigenetic regulation of *Tbx3* and *Prdm5*.

In conclusion, our study describes a mechanism of lineage determination in the development of primary liver cancer. Primary human liver carcinomas almost inevitably develop in chronically damaged livers in which different types of cell death such as necrosis, apoptosis or necroptosis occur. Our data suggest that hepatocytes with aberrantly activated oncogenes give rise to cholangiocarcinoma, when embedded in a necroptosis-dominated hepatic microenvironment. However, a hepatocyte that harbours the same oncogenic driver will give rise to HCC if it is not adjacent to necroptotically dying hepatocytes (Fig. 5h).

The necroptosis microenvironment is characterized by specific cytokines and our functional experiments suggest that these cytokines are secreted from immune cells that are activated by damage-associated molecular patterns released from necroptotically dying hepatocytes. Future work is needed to determine which cytokines act on oncogenically transforming hepatocytes and which intracellular signalling cascades in hepatocytes mediate the epigenetic regulation of *Tbx3* and *Prdm5* expression in ICC and HCC.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0519-y>.

Received: 28 August 2017; Accepted: 4 August 2018;

Published online 12 September 2018.

- Farazi, P. A. & DePinho, R. A. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nat. Rev. Cancer* **6**, 674–687 (2006).
- Rizvi, S. & Gores, G. J. Pathogenesis, diagnosis, and management of cholangiocarcinoma. *Gastroenterology* **145**, 1215–1229 (2013).
- Sirica, A. E. & Gores, G. J. Desmoplastic stroma and cholangiocarcinoma: clinical implications and therapeutic targeting. *Hepatology* **59**, 2397–2402 (2014).
- Wu, W. et al. Pattern of distant extrahepatic metastases in primary liver cancer: a SEER based study. *J. Cancer* **8**, 2312–2318 (2017).
- Fan, B. et al. Cholangiocarcinomas can originate from hepatocytes in mice. *J. Clin. Invest.* **122**, 2911–2915 (2012).
- Li, X. et al. Co-activation of PIK3CA and Yap promotes development of hepatocellular and cholangiocellular tumors in mouse and human liver. *Oncotarget* **6**, 10102–10115 (2015).
- Nkontchou, G. et al. Peripheral intrahepatic cholangiocarcinoma occurring in patients without cirrhosis or chronic bile duct diseases: epidemiology and histopathology of distant nontumoral liver in 57 White patients. *Eur. J. Gastroenterol. Hepatol.* **25**, 94–98 (2013).
- Schulz, P. O. et al. Association of nonalcoholic fatty liver disease and liver cancer. *World J. Gastroenterol.* **21**, 913–918 (2015).

9. Kang, T. W. et al. Senescence surveillance of pre-malignant hepatocytes limits liver cancer development. *Nature* **479**, 547–551 (2011).
10. Dauch, D. et al. A MYC-aurora kinase A protein complex represents an actionable drug target in p53-altered liver cancer. *Nat. Med.* **22**, 744–753 (2016).
11. Zender, L. et al. Cancer gene discovery in hepatocellular carcinoma. *J. Hepatol.* **52**, 921–929 (2010).
12. Gürlevik, E. et al. Adjuvant gemcitabine therapy improves survival in a locally induced, R0-resectable model of metastatic intrahepatic cholangiocarcinoma. *Hepatology* **58**, 1031–1041 (2013).
13. Marquardt, J. U., Andersen, J. B. & Thorgeirsson, S. S. Functional and genetic deconstruction of the cellular origin in liver cancer. *Nat. Rev. Cancer* **15**, 653–667 (2015).
14. Iverson, S. V., Comstock, K. M., Kundert, J. A. & Schmidt, E. E. Contributions of new hepatocyte lineages to liver growth, maintenance, and regeneration in mice. *Hepatology* **54**, 655–663 (2011).
15. Fujita, T. & Narumiya, S. Roles of hepatic stellate cells in liver inflammation: a new perspective. *Inflamm. Regen.* **36**, 1 (2016).
16. Grasl-Kraupp, B. et al. In situ detection of fragmented DNA (TUNEL assay) fails to discriminate among apoptosis, necrosis, and autolytic cell death: a cautionary note. *Hepatology* **21**, 1465–1468 (1995).
17. Luedde, T., Kaplowitz, N. & Schwabe, R. F. Cell death and cell death responses in liver disease: mechanisms and clinical relevance. *Gastroenterology* **147**, 765–783.e4 (2014).
18. Gautheron, J. et al. A positive feedback loop between RIP3 and JNK controls non-alcoholic steatohepatitis. *EMBO Mol. Med.* **6**, 1062–1074 (2014).
19. Afonso, M. B. et al. Activation of necroptosis in human and experimental cholestasis. *Cell Death Dis.* **7**, e2390 (2016).
20. Pasparakis, M. & Vandenabeele, P. Necroptosis and its role in inflammation. *Nature* **517**, 311–320 (2015).
21. Ofengeim, D. & Yuan, J. Regulation of RIP1 kinase signalling at the crossroads of inflammation and cell death. *Nat. Rev. Mol. Cell Biol.* **14**, 727–736 (2013).
22. Murphy, J. M. et al. The pseudokinase MLKL mediates necroptosis via a molecular switch mechanism. *Immunity* **39**, 443–453 (2013).
23. Postic, C. et al. Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. *J. Biol. Chem.* **274**, 305–315 (1999).
24. Chaisaingmongkol, J. et al. Common molecular subtypes among Asian hepatocellular carcinoma and cholangiocarcinoma. *Cancer Cell* **32**, 57–70.e3 (2017).
25. Conrad, M. L. et al. Maternal TLR signaling is required for prenatal asthma protection by the nonpathogenic microbe *Acinetobacter lwoffii* F78. *J. Exp. Med.* **206**, 2869–2877 (2009).
26. Cheng, H. Y., Chen, X. W., Cheng, L., Liu, Y. D. & Lou, G. DNA methylation and carcinogenesis of PRDM5 in cervical cancer. *J. Cancer Res. Clin. Oncol.* **136**, 1821–1825 (2010).
27. Suzuki, A., Sekiya, S., Büscher, D., Izpisua Belmonte, J. C. & Taniguchi, H. Tbx3 controls the fate of hepatic progenitor cells in liver development by suppressing p19ARF expression. *Development* **135**, 1589–1595 (2008).

Acknowledgements We thank E. Rist, P. Schieman, C. Fellmeth, C.-J. Hsieh, D. Heide and J. Hetzer for technical help or assistance. We thank A. Weber

for providing TLR2 and TLR4 knockout mice and W. S. Alexander and The Walter and Eliza Hall Institute of Medical Research for providing *Mikl^{fl/fl}* mice. The Cas9n-p19^{Arf} sgRNA vector was provided by W. Xue. We thank the c.ATG facility of Tuebingen University and CeGaT Tuebingen for exome sequencing and data analysis. This work was supported by the ERC Consolidator Grant ‘CholangioConcept’ (to L.Z.), the German Research Foundation (DFG): grants FOR2314, SFB685, SFB/TR209 and the Gottfried Wilhelm Leibniz Program (to L.Z.). Further funding was provided by the German Ministry for Education and Research (BMBF) (e:Med/Multiscale HCC), the German Universities Excellence Initiative (third funding line: ‘future concept’), the German Center for Translational Cancer Research (DKTK), the German-Israeli Cooperation in Cancer Research (DKFZ-MOST) (to L.Z.) and the Intramural Research Program of the Centre for Cancer Research, National Cancer Institute, National Institutes of Health (to X.W.W.). The group of O.B. is supported by grants from ANR-BMFT, Fondation ARC pour la recherche sur le Cancer, INSERM, and the National Cancer Institute of the National Institutes of Health under Award Number R01CA136533. O.B. is a CNRS fellow.

Reviewer information Nature thanks E. Guccione, E. Pikarsky and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The study was designed by L.Z., M.S. and F.H. with support from O.B. T.B. provided TLR KO (TLR2, 3, 4, 7 and 9 KO) mice. Mouse experiments, western blots, qRT-PCR, immunohistochemistry, immunofluorescence, vector generation and cell culture work were conducted and analysed by M.S., F.H. and L.Z. L.D. performed and analysed flow cytometry experiments, J.H. performed immunohistochemistry, immunofluorescence and mouse sampling, L.H. crossed *ROSA^{mT/mG} × Alb-cre × p19Arf^{-/-}* mice, S.K. and T.-W.K. conducted mouse experiments, R.C. subcloned vectors and performed knockdown experiments. Histopathological analyses were performed by T.Lo. and B.S. Human ICC and HCC samples were collected and analysed by H.D. and X.W.W. ChIP-seq, ATAC-seq, transcriptome and integrative analyses were performed by P.-F.R., O.B., G.D., N.R., L.R., M.R., J.Z. and M.H. M.V. and T.Lu. generated the *Alb-cre × Mikl^{fl/fl}* mice and conducted MLKL western blot analyses. L.Z. supervised the overall execution of experiments and analysed data. The manuscript was written by L.Z. with support from, M.S., F.H. and O.B.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0519-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0519-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Vector design. Sleeping beauty transposase (SB13), *NRAS*^{G12V} and *Myc* transposon plasmids have been described recently^{18–20}. The human myristoylated *Akt1* cDNA was provided by S. W. Lowe. The *Myc* and *Akt1* transposon plasmid (pCaMIA) was generated by PCR cloning of *Akt1* cDNA, introducing NcoI and SalI restriction sites, and subcloning into an MSCV vector. Then IRES-*Akt1* sequence was PCR amplified with primers introducing BspEI and AgeI and cloned into the pCaggs c-Myc transposon via AgeI. The *NRAS*^{G12V} cDNA was subcloned from pCaggs *NRAS*^{G12V} into an MSCV plasmid with BspHI and SalI restriction sites. IRES-*NRAS*^{G12V} was shuttled from MSCV into pCaggs c-Myc transposon plasmid via PCR cloning using primers with NotI and AgeI restriction sites to obtain pCaMIN. The full-length cDNA of *Tbx3* or *Prdm5* was purchased from Biocat (clone BC096551-TCM1004-GVO-TRI and BC138901-TCM1004-GVO-TRI) and shuttled via PCR cloning using primers with AscI/AgeI or AscI/NheI restriction sites into an empty pCaggs transposon plasmid to obtain pCaggs-*Tbx3* or pCaggs-*Prdm5*, respectively.

Animal strains. Mice were maintained in a specific-pathogen-free environment according to the University Hospital of Tuebingen guidelines and fed with a standard diet. *p19Arf*^{-/-} mice have been generated by C. Sherr. They were obtained in a C57BL/6 background from S. W. Lowe. Wild-type mice (C57BL/6, CB17 and C3H/HeN) and immunodeficient SCID/beige as well as *Rag2*^{-/-} knockout mice were purchased from Charles River or the Jackson Laboratory. *Alb-cre* × *Mlkl*^{fl/fl} mice^{22,23} (C57BL/6 background) were provided by M. Vucur and T. Luedde. For establishment of the lineage-tracing mouse model, *Alb-cre* mice were purchased from The Jackson Laboratory and crossed with the *ROSA*^{mt/mG} mouse strain that was generated by L. Luo and obtained from J. Zuber. *Alb-cre* × *ROSA*^{mt/mG} mice¹⁴ were intercrossed with *p19Arf*^{-/-} mice. Knockout mice lacking TLR2, TLR3, TLR4, TLR7 and TLR9 (C57BL/6 background)²⁵ were provided by T. Buch. TLR2 and TLR4 knockout (C3H/HeN background) mice were obtained from A. Weber. Animal experiments were approved by the local authorities (Regierungspräsidium Tuebingen, Baden-Wuerttemberg, Germany). No tumours exceeded the approved tumour size of 0.5 cm.

Hydrodynamic tail vein injection and in vivo electroporation. Vectors for HDTV injection and Epo were prepared using the QIAGEN EndoFree Maxi Kit (QIAGEN). For transposon-mediated gene transfer, animals received a 25 µg:5 µg ratio of transposon to transposase-encoding plasmid. For HDTV, DNA was suspended in 0.9% saline solution at a final volume of 10% of the animal's body weight and injected via the tail vein within 5–10 s.

For the electroporation of the left lateral liver lobe, mice were anaesthetized using ketamine (100 mg kg⁻¹ body weight) and xylazine (10 mg kg⁻¹ body weight) via intraperitoneal injection. A small midline laparotomy was done and the liver was atraumatically luxated. The left liver lobe was injected with DNA solution in a total volume of 50 µl TE buffer with a 27G needle. Electroporation was performed with the Square Wave Electroporator (CUY21SC, Nepa Gene) using an electric pulse via a tweezer-type electrode (CUY650P5, 5 mm diameter). Electric pulses were applied twice with a duration of 75 ms, an interval of 500 ms and a voltage of 75 V.

Bile duct ligation. For the ligation of the common bile duct, mice were anaesthetized using ketamine (100 mg kg⁻¹ body weight) and xylazine (10 mg kg⁻¹ body weight) via intraperitoneal injection. A small midline laparotomy was made and the liver lobes were atraumatically luxated and fixed with a fleece compress moistened with sterile PBS. The bile ducts were carefully separated from the portal vein and hepatic artery using a micro-serrations forceps. The common bile duct was ligated with two surgical knots.

Liver perfusion. Mouse hepatocytes were isolated via liver perfusion through the vena cava with liver perfusion medium (Invitrogen) for 15 min and then with collagenase (Serva) and Ca²⁺-supplemented medium for the next 15 min.

In vivo treatments. Mice were treated with Nec-1 (Cayman Chemical) at a dose of 10 µg per 100 µl in 0.9% NaCl three times per week intraperitoneally, starting 3 days before electroporation.

Subcutaneous tumour model. Cells were cultured with 10% fetal calf serum (FCS) in DMEM medium at 37 °C and 7% CO₂. For isolation and cultivation of primary tumour cells the tumour was dissected under sterile conditions and minced with a scalpel in DMEM, 1 × HBS, dispase (1,000 U ml⁻¹) and collagenase (0.1 U ml⁻¹) (Roche). Cells were incubated at 37 °C for 30 min and filtrated through a 100 µm nylon mesh. Cells were washed twice with PBS and seeded on gelatine (1%)-coated plates. For immunofluorescence of K19 and HNF4α, cells were plated on chambers and incubated at 37 °C and 7% CO₂ overnight. Cells were washed twice with PBS, fixed for 10 min with 4% paraformaldehyde (PFA) and subsequently stained using standard protocols with primary antibodies K19 (1:100, TROMAIII, DSHB) and the secondary antibodies (1:1,000, Invitrogen A11032) and mounted with H-1500 Vectashield Hardset with DAPI mounting medium. For subcutaneous

injections, cultured tumour cells were washed twice with sterile PBS, trypsinized, resuspended in DMEM and filtrated through a 100 µm nylon mesh. Cells were counted and 2 × 10⁶ cells in 100 µl PBS were injected in the left or right rear flank of immunodeficient *Rag2*^{-/-} mice.

Generation of stable knockdown cell lines. Phoenix-AMPHO cells were transfected via standard calcium phosphate method with doxycycline-inducible retrovirus plasmids harbouring shRNA cassettes, puromycin selection cassettes and GFP marker gene. After 2 days, viral supernatant was collected and filtered with 0.45 µm sterile filters. Target cells were treated with polybrene (10 µg ml⁻¹) and infected with 100 µl viral supernatant. Efficiently infected cells were selected with Puromycin (4–6 µg ml⁻¹). Successful selection was checked after induction of the GFP and shRNA cassette with doxycycline (15 µg ml⁻¹) with a fluorescence microscope (Olympus CKx41).

Western blot. Whole-cell extracts were prepared using RIPA buffer and protein concentration was measured with the BioRad DC assay. 25 µg of protein were subjected to 15% SDS-PAGE and blotted on a PVDF membrane (Millipore) with a semi-dry blot or wet blot system. The following antibodies were used: cleaved caspase 3 (Cell Signal, 1:1,000), pMLKL (S345) (abcam, 1:1,000), vinculin (Thermo Fischer 1:10,000), tubulin (Cell Signal, 1:10,000) and β-actin (Sigma, 1:10,000). Blots were visualized using Super Signal West Kit (Thermo Scientific) and the ChemiDocTM MP Imaging System. Freshly isolated hepatocytes were used to show absence of MLKL in *Alb-cre* × *Mlkl*^{fl/fl} mice. Primary hepatocytes were homogenized in NP-40 lysis buffer (50 mM Tris-HCL (pH 7.5), 150 mM NaCl, 0.5% NP-40 supplemented with PhosSTOP phosphatase inhibitor (Roche), cComplete protease inhibitor (Roche), 1 mM Pefabloc (Roche) and 1 mM 1,4-dithiothreitol (DTT, Roth)) buffer to gain protein lysates. MLKL (Biorbyt orb32399, 1:2,000) antibody was used to detect MLKL in samples. If necessary, stripping has been performed with Restore Western Blot Stripping Buffer (Biorad).

RNA isolation and qRT-PCR. Total RNA was isolated from liver tissue and freshly cultured cells using the TRIZOL RNA isolation protocol from Invitrogen and the RNeasy Kit (Qiagen). qPCR was performed with an ABI 7300 Real-Time PCR System (Applied Biosystems) in duplicates or triplicates. The TaqMan Reverse Transcription Reagents Kit (Invitrogen) was used for cDNA synthesis from total RNA. Supplementary Table 2 depicts primer sets for the detection of single genes. The qPCR analysis was carried out using the SYBR Green Master Mix from Applied Biosciences. Values were normalized towards *Actb* quantification. Cytokine profiling for pathway-focused gene expression analysis was performed using the mouse Inflammatory Cytokines and Receptors RT² Profiler PCR Array from Qiagen. Gene expression analysis for epigenetic chromatin modification enzymes was performed with the mouse RT² Profiler PCR Array Mouse Epigenetic Chromatin Modification Enzymes Array from Qiagen. RNA isolation and cDNA synthesis was performed as described above. The RT² Profiler PCR Array was used according to the manufacturer's protocol.

Flow cytometry analysis. The liver was chopped into small ~1 mm³ pieces and then enzymatically digested in a medium composed of equal volume of DMEM and HBS supplemented with 0.5 mg ml⁻¹ Collagenase (Serva Collagenase NB 4G) for 30 min at 37 °C. The enzymatic reaction was stopped using cold medium and the liver suspension was meshed through a 70 µm nylon mesh (Falcon). After centrifugation erythrocytes were lysed using an ACK buffer (150 mM NH₄Cl, 10 mM KHCO₃ and 0.1 mM EDTA). 1 million of cells were resuspended in blocking solution (2% BSA in PBS) and stained with antibodies (Supplementary Table 3) on ice for 30 min. Samples were immediately acquired using a FACSCanto flow cytometer (BD Biosciences) from the flow cytometry core facility Tuebingen. Doublets were excluded using height versus area dot plots and samples were additionally gated on viable leukocytes by DAPI exclusion in Extended Data Fig. 7b. Detailed gating strategies are provided in the flow cytometry Reporting Summary. Data analysis was performed using FlowJo software (Tree Star).

Histopathology, immunohistochemistry and immunofluorescence of paraffin samples. Histopathological evaluation of mouse livers was performed by experienced board-certified pathologists (T.L. and B.S.) with H&E, HNF4α- and K19-stained paraffin-embedded liver tumour sections. For H&E, liver samples were fixed overnight in 4% PFA, paraffin-embedded, sectioned to 4-µm thickness, and subsequently used for haematoxylin and eosin following standard protocols. For RIPK3, pRIPK3, K19 and HNF4α staining of deparaffinized and hydrated sections, a heat-induced antigen retrieval method was performed for 10 min using sodium citrate buffer. Either liquid DAB plus substrate reagent (Zytomed) or the Metal Enhanced DAB Substrate Kit (Thermo Fischer) was used to perform direct chromogenic visualization. Primary antibodies used include K19 (1:100, TROMAIII, DSHB), pRIPK3 (5 µg ml⁻¹, Genentech), RIPK3 (0.5 µg ml⁻¹, Genentech) and HNF4α (1:100, Santa Cruz, sc-6556). Secondary antibodies used were biotinylated anti-rat (1:100, Vector Laboratories), biotinylated anti polyvalent (Thermo Scientific), biotinylated anti-goat (1:100, Santa Cruz) or Alexa-Fluor 594 labelled anti-rat (1:1,000, Invitrogen). Counting of HNF4α-positive cells was performed via an ImageJ script.

Immunostaining analysis of Ly6G, MHCII, B220 and CD3 and α SMA was performed on a BOND-MAX immunohistochemistry robot (Leica Biosystems) using BOND polymer refine detection solution for DAB with the following antibodies: Ly6G (1:600, BD Pharmingen, 1A8), MHCII (1:500, Novus Biologicals, M5/114.15.2), B220 (1:3,000, BD Pharmingen, RA3-6B2), CD3 (1:250, Zytomed Systems, SP7) and α SMA (Sigma A2547 1:5,000). Image acquisition was performed with a Leica SCN400 slide scanner or Olympus BX63 microscope and a DP80 camera (Olympus). Quantification of immune cells was performed via calculating DAB positive pixels per area and counted via an ImageJ script.

Native fluorescence detection. For the detection of native fluorescence, liver tissue was fixed for 4 h in 4% PFA at 4°C. The fixative solution was replaced by 30% sucrose (Roth) in PBS for overnight incubation at 4°C. Afterwards, the tissue was embedded in Tissue-Tec OCT compound (Sakura), frozen at -80°C, cut in 6- μ m thick tissue sections with an HM 560 cryostat and mounted with H-1500 Vectashield Hardset with DAPI mounting medium for native fluorescence expression analysis. Microscopic analysis was performed with the Olympus BX63 microscope and a DP80 camera (Olympus).

Cell death detection. For TUNEL testing (TdT-mediated dUTP nick end labelling) In situ Cell Death Detection Kit Fluorescein from Roche Diagnostics was used according to the manufacturer's instructions. For co-staining of TUNEL and native GFP in *ROSA^{mT/mG} × Alb-cre × p19^{Arf/-}* mice, the In situ Cell Death Detection Kit TMR-Red was used. The protocol was modified to preserve GFP signal but destroy the tomato signal. The fixation step was reduced to 5 min, the permeabilization step was reduced to 2 min, and incubation duration was reduced to 15 min. Sections were analysed with an Olympus BX63 and DP80.

Laser capture microdissection. Cryosections (6 μ m) on MembraneSlide 1.0 PEN (Zeiss) were stained with the Arcturus HistoGene LCM Frozen Section Staining Kit (Thermo Scientific) according to the manufacturer's protocol. Laser capture microdissection was performed with a Zeiss PALM MicroBeam and PALMRobo Software.

Isolation of genomic DNA. Tumour tissues were mixed with 300 μ l solution A (20 mM Tris-HCl, pH 8, 100 mM EDTA, 100 mM NaCl, 1% SDS and 0.5 mg ml⁻¹ proteinase K) minced with a homogenizer and incubated at 56°C overnight. Proteinase K was inactivated at 95°C for 5 min, 50 μ l 5 M NaCl was added and incubated again at 95°C for 5 min. Samples were centrifuged and supernatant was mixed with two-third volumes of isopropanol. After another centrifugation, the DNA pellet was washed with 70% ethanol and the dried pellet was resuspended in water.

Mutagenesis. For subcloning of *Fam72a* cDNA into the pCaggs transposon plasmid, cDNA (*Fam72a* MR201041, Mouse cDNA ORF Clone) was purchased from Biotac and subcloned using *AscI*/*AgeI* restriction sites into pCaggs to obtain pCaggs-*Fam72a*. Site-directed mutagenesis of *Fam72a* in transposon vector was performed using Q5 Site-Directed Mutagenesis Kit (NEB). Specific primers were designed to generate specific mutation at position 259 and obtain G>T mutation. The protocol was performed according to the manufacturer's manual.

Whole-exome sequencing. Whole DNA purification out of laser capture microdissection samples was performed with the QIAamp MicroKit (Qiagen) according to manufacturer's protocol. DNA concentration was measured with Qubit dsDNA HS Assay Kit (Thermo Scientific). After fragmentation with an ultrasonicator (Covaris), library preparation was performed via EndRepair and A-Tailing with KAPA Hyper Library Prep Kit (PeqLab), followed by Adaptor Ligation with Agilent SureSelect Oligo-Mix and purification with AMPureXP according to the protocol. PreCapPCR was performed according to the KAPA Hyper Library Prep Kit and purified with AMPureXP. Enrichment was done with the SureSelect Mouse AllExon XT Target Enrichment Kit (Agilent). Sequencing was done with a RapidRun on a HiSeq2500. Filtering of data was performed using SeqPurge, mapping was done with the BWA mem algorithm. Quality filtering was performed with MarkDuplicates from Picard tools. Alignment was done with BamLeftAlign and local realignment was performed as described elsewhere²⁸. Variant calling was done with Freebays and SnpEff and SnpSift were used for annotation and filtering of variant data.

ATAC-seq library preparation. ATAC was performed as described previously²⁹. In brief, 50,000 cells were harvested and centrifuged (500g, 5 min, 4°C). Cells were washed in cold PBS and centrifuged (500g, 5 min, 4°C). Nuclei were isolated by resuspension of the cell pellet in lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂, 0.1% IGEPAL CA-630) and centrifugation (500g, 5 min, 4°C). Nuclei were subjected to tagmentation in 50 μ l reactions containing 1 × TD buffer and 2.5 μ l T5 transposase followed by incubation at 37°C for 30 min. Tagmented DNA was recovered using a QIAGEN MinElute kit and eluted in 10 μ l elution buffer (10 mM Tris-HCl, pH 8.0). Libraries were amplified for 14 cycles, purified using a QIAGEN PCR Cleanup kit and eluted in 20 μ l elution buffer. Size distribution of libraries was determined on a Bioanalyzer 2100 instrument.

Microarray gene expression profiling. Total RNA was preprocessed for hybridization to Mouse Gene 2.0 ST Array (Affymetrix) using the GeneChip WT PLUS Reagent Kit (Affymetrix) following the manufacturer's protocol. In brief, 100 ng

of total RNA was used to generate first strand cDNA using reverse transcriptase and primers containing a T7 promoter sequence. Single-stranded cDNA was then converted to double-stranded cDNA by using DNA polymerase and RNase H to simultaneously degrade the RNA and synthesize second-stranded cDNA. Complementary RNA (cRNA) was synthesized and amplified by in vitro transcription (IVT) of the second-stranded cDNA template using T7 RNA polymerase. Subsequently, single-stranded cDNA was synthesized by the reverse transcription of cRNA with incorporated deoxyuridine triphosphate (dUTP). Purified, single-strand cDNA was fragmented by uracil-DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE 1) at the unnatural dUTP residues and labelled by terminal deoxynucleotidyl transferase (TdT) using the Affymetrix proprietary DNA Labelling Reagent that is covalently linked to biotin. Subsequent hybridization, wash, and staining were carried out using the Affymetrix GeneChip Hybridization, Wash, and Stain Kit following the manufacturer's protocols. In brief, each fragmented and labelled single-strand cDNA target sample (approximately 3.5 μ g) was individually hybridized to a GeneChip Mouse Gene 2.0 ST Array at 45°C for 16 h in Affymetrix GeneChip Hybridization Oven 640. After hybridization, the array chips were stained and washed using an Affymetrix Fluidics Station 450. The chips were then scanned on Affymetrix GeneChip Scanner 3000 7G.

ATAC-seq data processing. Paired-ends reads were cropped to 100 bp with trimomatic v0.36³⁰ and cleaned using cutadapt v1.8.3 to remove Nextera adapters, low quality bases and reads, and discard reads shorter than 25 bp after trimming. Fragments were then aligned to the mouse reference genome (mm10) with bowtie2 v2.2.3³¹ discarding inconsistent pairs and considering a maximum insert size of 2kb (bowtie2 -N 0 -no-mixed -no-discordant -minins 30 -maxins 2000). Alignment files were further processed with samtools v1.2 and PicardTools v1.130 to flag PCR and optical duplicates and remove alignments located in Encode blacklisted regions. Accessible regions were identified using MACS2 v2.1.0 in paired-end mode without control using default parameters. After assessing library saturation using preseqR, alignment and peak data were imported and pre-processed in R using the DiffBind³² package. We first defined the global peak set as the union of all eight peak sets defined previously. We then counted the number of reads mapping inside each of these intervals for each sample. The raw count matrix was then normalized for sequencing depth using a nonlinear full quantile normalization as implemented in the EDASeq³³ package. To remove sources of unwanted variation and consider batch effects, data were finally corrected with the RUVSeq³⁴ package considering one surrogate variable. Differential analyses for count data was performed using edgeR³⁵ considering the tumour type and surrogate variable in the design matrix, by fitting a negative binomial generalized log-linear model to the read counts for each peak. Peaks were finally annotated using ChIPpeakAnno considering annotations provided by Ensembl v86. Reads assigned per million map reads (RPM) normalized visualization tracks were generated using deepTools³⁶.

ChIP-seq library preparation. HCC (H1 and H4), and ICC (E9 and E10) cell lines were grown in DMEM high glucose glutamax (GIBCO) media in 600 cm² tissue culture dishes to 80% confluence. Approximately 50 million cells were collected in 10 million cell aliquots in 15 ml media. Each aliquot was cross-linked in 1% paraformaldehyde for 10 min at room temperature. Cross-linking was quenched with the addition of 1 M of 2 M glycine, and incubated at room temperature for 5 min. Chromatin was isolated and digested with 1.2 μ l of micrococcal nuclease using the Cell Signalling SimpleChIP kit (9002). Approximately 40 million cell equivalents of pre-cleared chromatin were used to perform immunoprecipitation following the cell signalling protocol. Immunoprecipitations for TBX3 with HCC H1 and H4 chromatin were performed with an anti-TBX3 antibody (Santa Cruz, sc-17871, 5 μ g), and the immunoprecipitations for PRDM5 with ICC E9 and E10 chromatin were performed with anti-PRDM5 antibody (Millipore, MABE972, 5 μ g). Inputs were derived from 500,000 cell equivalents of chromatin.

ChIP-seq libraries were produced using a modified protocol from the Accel-NGS 2S Plus DNA Library Kit (21024), in which we performed DNA extraction using 25:24:1 phenol:chloroform:isoamyl alcohol followed by overnight ethanol precipitation of DNA at each step of the protocol. In addition, we followed an enrichment protocol for small DNA fragments outlined in the X-ChIP protocol³⁷. ChIP-seq libraries underwent quality control using the Agilent Technologies 4200 Tapestation (G2991-90001) and quantified using the Invitrogen Qbit DS DNA HS Assay kit (Q32854). Libraries were sequenced in single-end 65-bp on an Illumina HiSeq 2500, generating 262 million reads (33 million on average per sample).

ChIP-seq data processing. Single-ends 65-bp reads were cleaned using fastq-mcf v1.04.803 from the ea-utils suite³⁸ to remove Illumina adapters, low quality bases and reads, and discard reads shorter than 25 bp after trimming. Fragments were then aligned to the mouse reference genome (mm10) with bowtie v1.1.1³⁹ using best matches parameters (bowtie -v 2 -m 1 -best -strata). Alignment files were further processed with samtools v1.2⁴⁰ and PicardTools v1.130 to flag PCR and optical duplicates and remove alignments located in Encode blacklisted regions. For the follow-up analyses, we ended up with 214 million ready-to-analyse alignments (27 million on average per sample). Fragment size was estimated in

silico for each library using spp. v1.10.1. Genome-wide consistency between replicates was checked using custom R scripts. Enriched regions were identified for each replicate independently with MACS v2.1.0⁴¹ with non-immunoprecipitated genomic DNA as a control (macs2 callpeak –nomodel –shiftsize –shift-control –gsize mm –p 1e-1). These relaxed peak lists were then processed through the irreproducible discovery rate pipeline⁴² to generate an optimal and reproducible set of peaks for each transcription factor. Peaks were finally annotated using ChIPpeakAnno considering annotations provided by Ensembl v.86.

Normalized ATAC-seq and ChIP-seq signal tracks. The genome was binned in 200-bp non-overlapping windows and we generated genome-wide read count matrices for each assay or sample independently. Counts were finally transformed to reads per million mapped reads (RPM) and data were quantile normalized with custom R script to generate scaled and normalized signal tracks. These data were further used to build genome browser views and density heat maps.

Mouse microarray analyses. Raw Affymetrix Mouse Gene 2.0 ST array intensity data were analysed using open-source Bioconductor packages on R. The data were normalized all together (4 HCC and 4 ICC) using the robust multi-array average normalization approach implemented in the oligo package⁴³. Internal control probe sets were removed and average expression deciles over conditions were then defined. Probes in which the average expression was lower than the fourth expression decile were removed for subsequent analyses. To remove sources of unwanted variation and consider batch effects, data were finally corrected with the sva package. To get as much information as possible, we combined Affymetrix Mouse Gene 2.0 ST annotations provided by Affymetrix and Ensembl through the packages pd.mogene.2.0.st and biomaRt⁴⁴. Principal component analysis and bi-clustering based on Pearson's correlation and Ward's aggregation criterion were used to check for consistency between biological replicates and experimental conditions at each step of the pre-processing. Normalized log-scaled and filtered expression data were further considered for differential analysis with limma⁴⁵. In brief, a nested general linear model considering tumour type (HCC or ICC), gender (male or female) and their interaction was fitted to the data. Moderated *F*-statistics that combine the empirical Bayes moderated *t*-statistics for all contrasts into an overall test of significance for each probe were used to assess the significance of the observed expression changes when comparing HCC and ICC. *P* values were corrected for multiple testing using the false discovery rate (FDR) approach for a significance level of 0.1. A cut-off in fold-change at 1.5 was finally applied.

Mouse microarray analyses, shRNA experiments. Raw Affymetrix Mouse Gene 2.0 ST array intensity data were preprocessed altogether for the two shRNA experiments (2 shRNAs against *Prdm5* in ICC, 2 shRNAs against *Tbx3* in HCC, each with biological duplicates, and two biological duplicates for the shRNA control in each cell type, that is, 12 arrays), following the same procedure as described above. After differential analysis with limma to test for the effect of the two shRNAs independently, we split deregulated genes as either direct or indirect *Prdm5* or *Tbx3* targets according to the presence or absence of proximal ChIP-seq peaks (<100 kb from the TSS or inside the gene body of deregulated genes). We finally performed gene set over-representation analyses on global, direct and indirect up- and down-regulated genes after *Tbx3* or *Prdm5* knockout, considering the MSigDB canonical gene sets using a right-tail modified Fisher's exact test and the hypergeometric distribution to provide *P* value.

Microarray analysis of human cohorts. The microarray data from the 199 Thailand Initiative in Genomics and Expression Research for Liver Cancer (TIGER-LC) cohort of HCC and ICC can be found on Gene Expression Omnibus (GEO) accession GSE76297 (<http://www.ncbi.nlm.nih.gov/geo>). In brief, log₂ gene

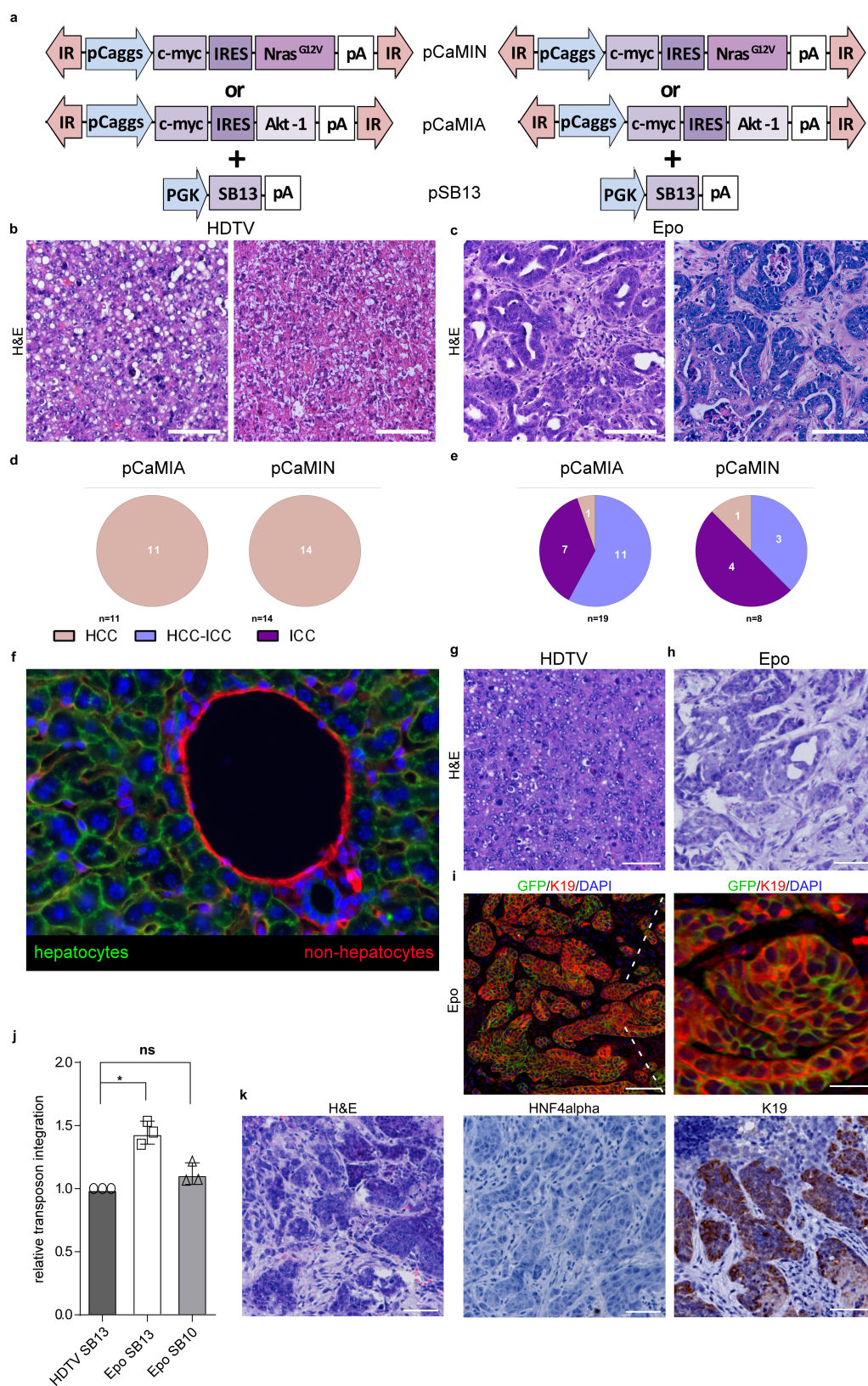
expression values of specified genes from tumour samples only were extracted for comparative analyses between HCC and ICC. Non-parametric statistical analyses using Mann–Whitney test were performed to test for differences amongst two groups using GraphPad Prism (v.7.01).

Statistics. GraphPad Prism (Version 6) was used for statistical analyses. Unpaired Student's *t*-test was used to calculate *P* values, unless stated otherwise in the figure legends.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The data and code that support the findings of this study are available from the corresponding author on reasonable request. Source data for graphs showed in Figs. 4, 5 and Extended Data Figs. 1, 4–10 are available in the online version of this paper. Data from ChIP-seq experiments are available at the Sequence Read Archive (SRA) under the accession number SRP136997. Whole scans of western blots are depicted in Supplementary Fig. 1, and the gating strategy for flow cytometry is depicted in Supplementary Fig. 2.

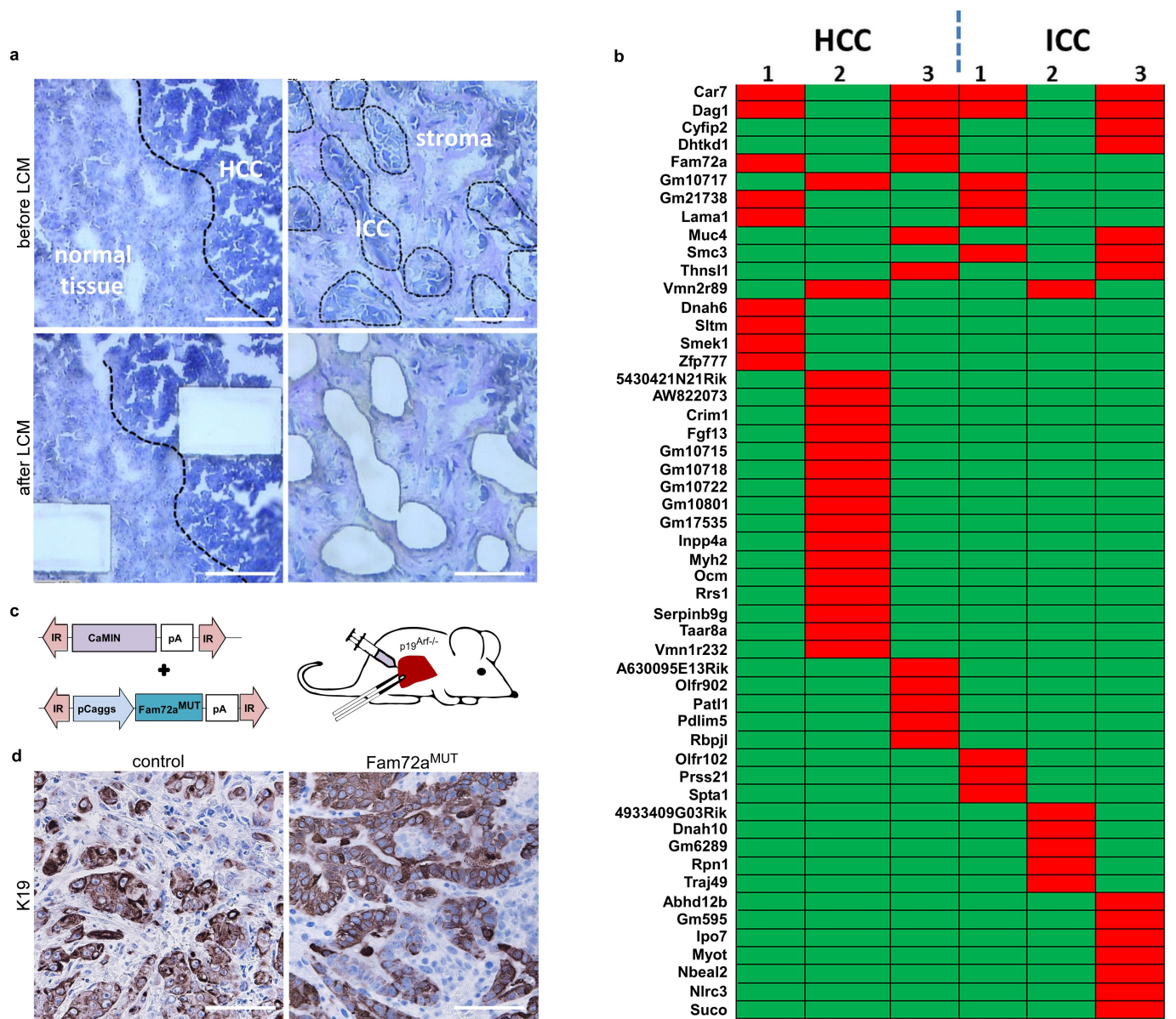
28. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
29. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
32. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
33. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
34. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
35. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
36. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
37. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* **4**, e09225 (2015).
38. Aronesty, E. Comparison of sequencing utility programs. *Open Bioinformatics* **7**, 1–8 (2013).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
40. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Feng, J. et al. Using MACS to identify peaks from ChIP-seq data. *Bioinformatics* **34**, 2.14.1–2.14.14 (2011).
42. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
43. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
44. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protocols* **4**, 1184–1191 (2009).
45. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).



Extended Data Fig. 1 | See next page for caption.

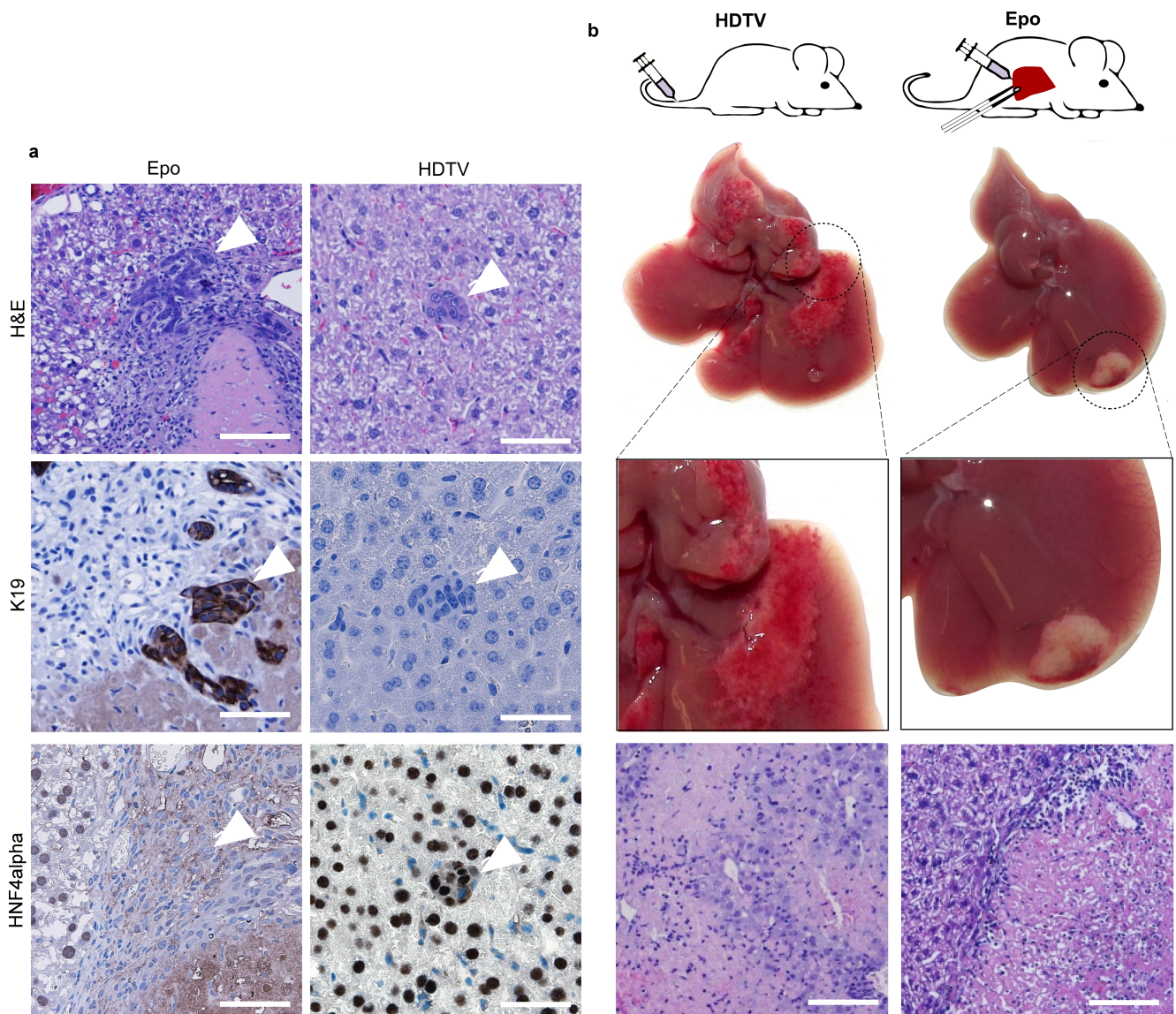
Extended Data Fig. 1 | Tumour phenotype depends on the delivery method of oncogene encoding transposons. **a**, Schematic representation of transposon vectors encoding *Myc* and *NRAS*^{G12V} (pCaMIN) or *Myc* and *Akt1* (pCaMIA) and a plasmid encoding the SB13 transposase. **b, c**, Representative micrographs of H&E staining of HDTV- or Epo-derived tumours. Scale bars, 100 μ m. **d**, Histopathological scoring and quantification of tumours developed after hydrodynamic delivery of oncogene encoding transposons. **e**, Histopathological scoring and quantification of tumours developed after transposon delivery via in vivo electroporation. **f**, Representative image of native fluorescence microscopy of liver cryosections from *ROSA*^{mT/mG} \times *Alb-cre* \times *p19*^{Arf-/-} mice. In such mice, activation of the albumin promoter induces excision of a red fluorescence marker gene (mTomato) together with a stop codon flanked by *loxP* sites, thus resulting in a colour switch from red to green fluorescence (membrane-bound GFP). In this model, only fully differentiated hepatocytes (with high albumin promoter activity and therefore high levels of Cre expression) were able to induce the switch from red to green fluorescence, whereas liver cells with low albumin promoter activity such as embryonic hepatocytes or oval cells or liver

progenitor cells were unable to accomplish such a colour change. Shown is mGFP expression in hepatocytes (green) and mTomato expression in bile duct cells or endothelial cells (red) ($n = 3$). Scale bar, 100 μ m. **g, h**, Representative H&E staining images of tumours 4 weeks after HDTV (**g**) or Epo (**h**) transfection of the pCaMIN vector in *ROSA*^{mT/mG} \times *Alb-cre* \times *p19*^{Arf-/-} mice ($n = 4$). Scale bars, 100 μ m. **i**, Representative images of DAPI-positive (blue), K19-positive (red) and native GFP-positive (green) hepatocytes in ICC derived from pCaMIN electroporated *ROSA*^{mT/mG} \times *Alb-cre* \times *p19*^{Arf-/-} mice ($n = 6$, left). Scale bars, 100 μ m (left) and 20 μ m (right). Data are from one experiment. **j**, qPCR analysis with transposon-specific primers on DNA isolated from HDTV- or Epo-induced tumours using (SB13) showed an approximately 1.5-fold increased transposon integration compared to tumours triggered by hydrodynamic delivery (HDTV). Epo-induced tumours using the SB10 transposase show equal transposon integration levels compared to HDTV-derived tumours with SB13 ($n = 3$). NS, not significant ($P = 0.074$); $*P = 0.0011$, Student's two-sided *t*-test. Data are mean \pm s.d. **k**, Representative images of H&E, K19 or HNF4 α staining of Epo-induced tumours transfected using pCaMIN and SB10 ($n = 3$). Scale bars, 100 μ m.



Extended Data Fig. 2 | Exome sequencing reveals recurrent mutations in HCC and ICC. **a**, Purification of epithelial components from HCC or ICC derived from pCaMIN electroporated $p19^{Arf-/-}$ mice and normal liver tissue as a control using laser capture microdissection (LCM) ($n = 3$ per group). Scale bars, 100 μm . **b**, Exome sequencing revealed recurrent mutations (in red), in which 12 mutations were found in at least 2 samples

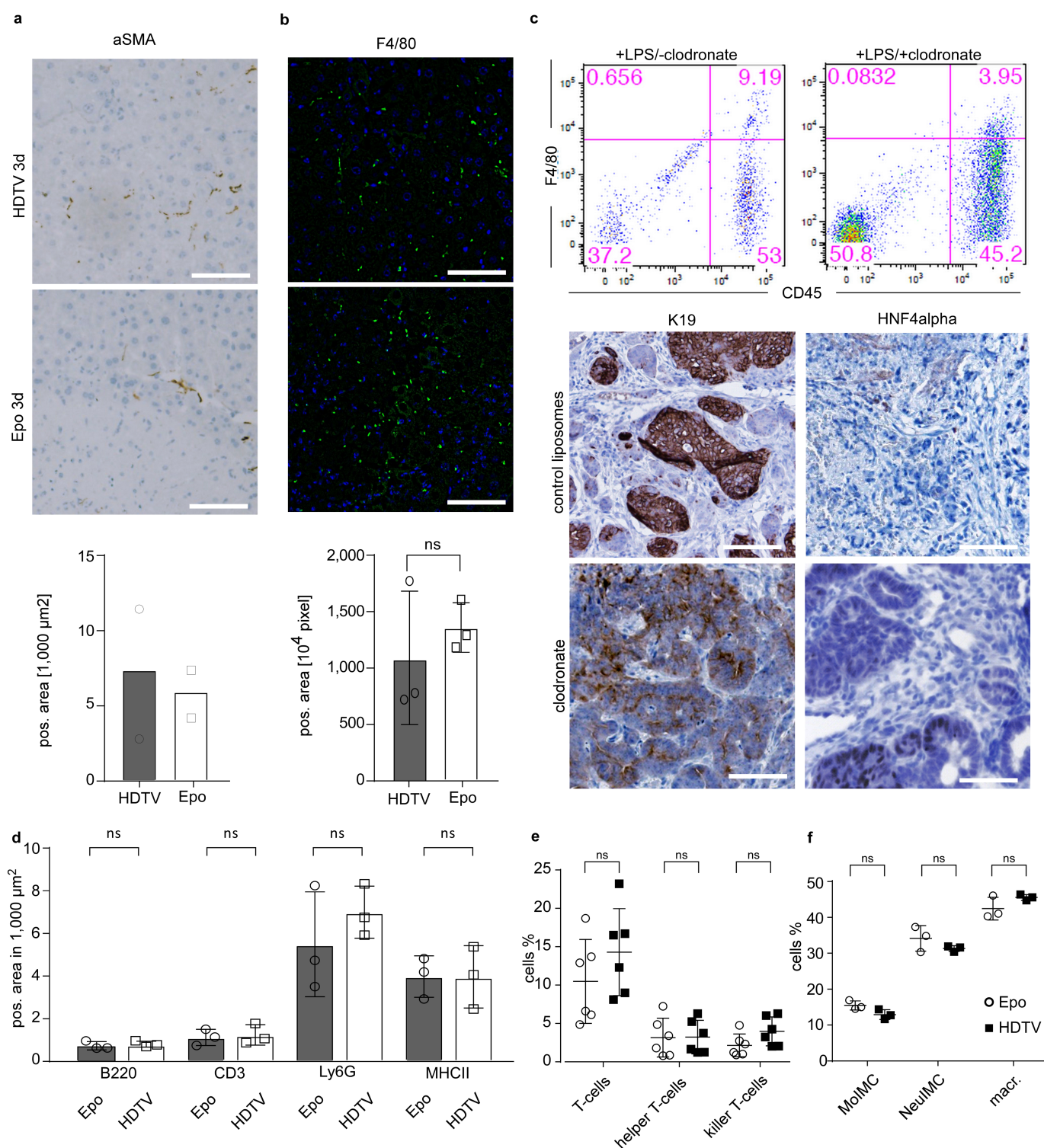
in 3 analysed HCC (left) and 3 ICC (right) tissues. **c**, Schematic outline of transposon vectors expressing *Myc* and *NRAS*^{G12V} (pCaMIN) and mutated (259G>T) *Fam72a* cDNA (bottom), which were co-delivered into $p19^{Arf-/-}$ mice. **d**, Immunohistochemical analysis of tumour tissue for K19 expression ($n = 3$ per group). Scale bar, 100 μm .



Extended Data Fig. 3 | Characterization of early and pre-tumorigenic phase after Epo- or HDTV-mediated oncogene delivery.

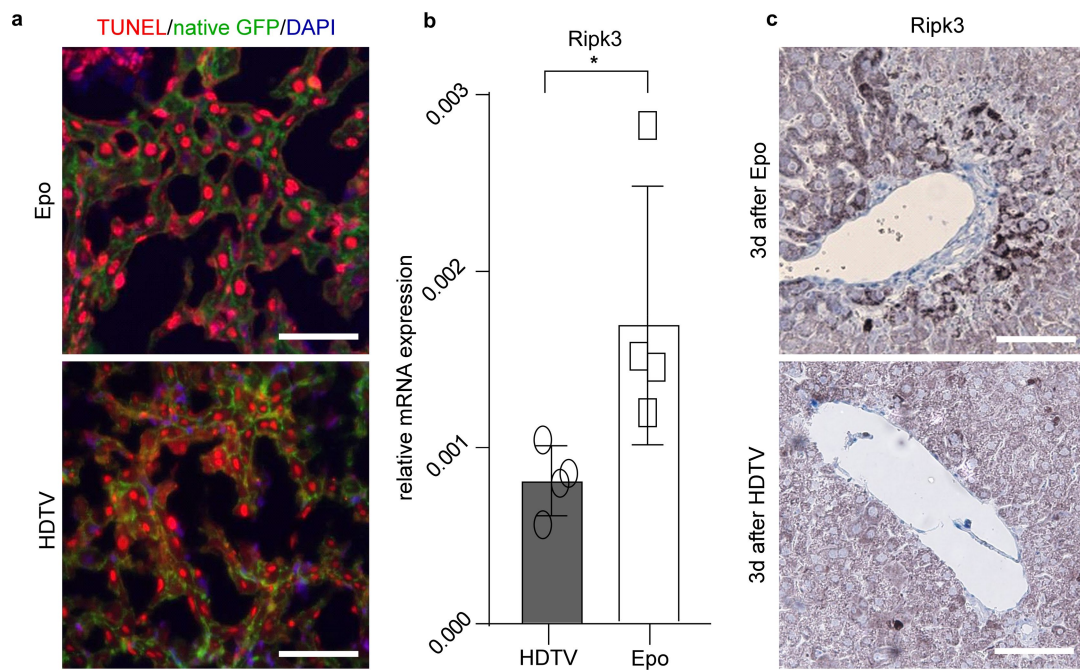
a, Immunohistochemical analysis of $p19^{Arf-/-}$ deficient liver sections 5 days after Epo- or HDTV-mediated transposon delivery, showing microtumours in H&E (top) and Epo-derived K19-positive, HNF4 α -negative ICCs (middle and bottom left panel) as well as HDTV-derived HNF4 α -positive, K19-negative HCCs (middle and bottom right panel, indicated

by white arrowheads) ($n = 3$). Scale bars, 100 μm . **b**, Schematic outline of the experimental approach (left) and representative macroscopic liver photographs 3 days after hydrodynamic (HDTV) or Epo delivery of the pCaMIN and SB13 vectors into $p19^{Arf-/-}$ mouse livers. Macroscopically visible liver damage (left) as well as eosinophilic areas indicating microscopic liver damage (right) are shown on H&E-stained liver sections ($n = 4$). Original magnification, $\times 200$.



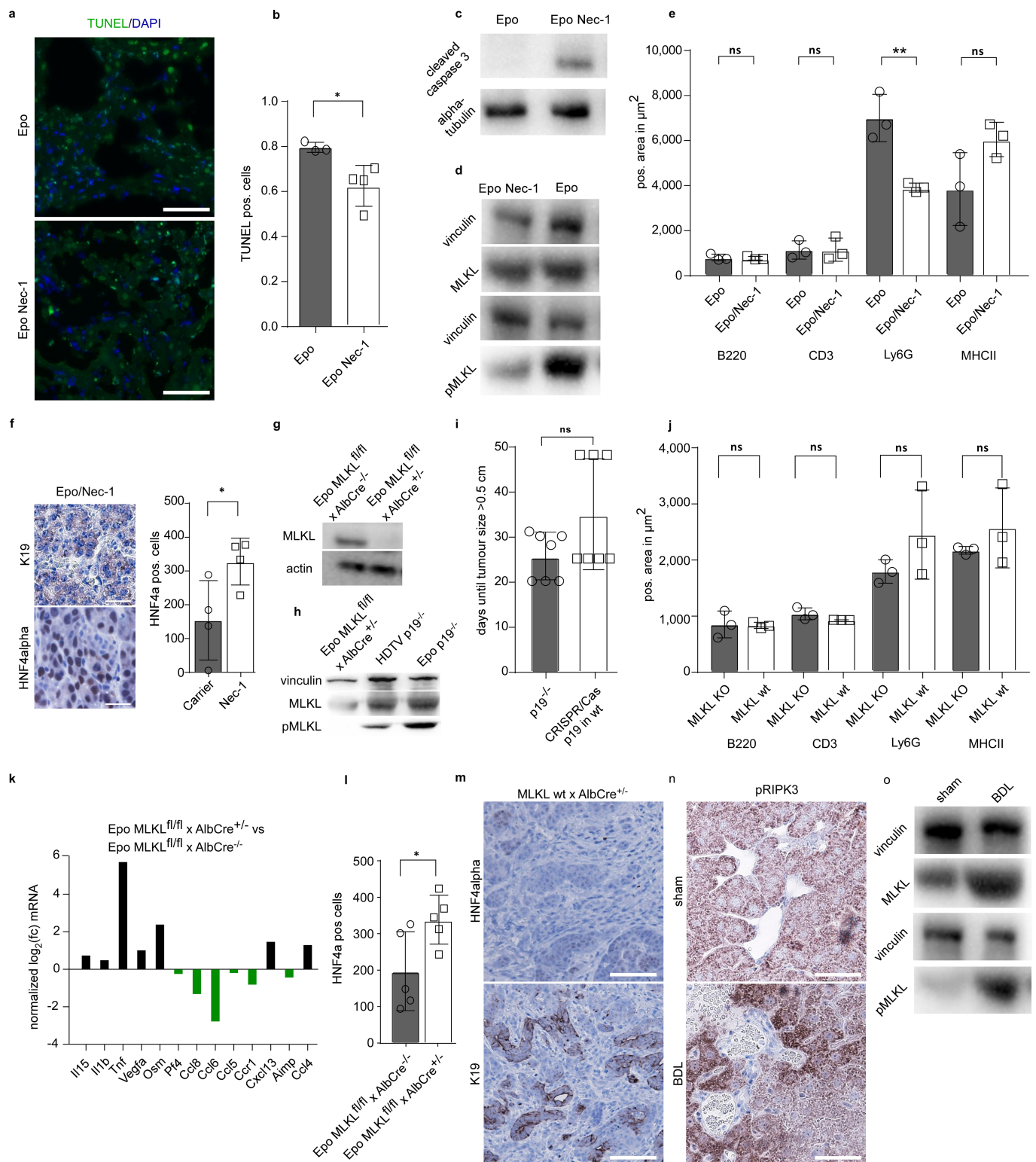
Extended Data Fig. 4 | Immune composition does not contribute to lineage commitment in liver cancer. **a**, Representative micrographs of α SMA immunohistochemistry (top) and quantification (bottom) 3 days after Epo and HDTV treatment in $p19^{Arf-/-}$ livers and quantification ($n=2$). Scale bars, 100 μ m. Data are mean \pm s.d. **b**, Representative micrographs of F4/80 immunofluorescence (top) and quantification (bottom) 3 days after Epo and HDTV treatment in $p19^{Arf-/-}$ livers ($n=3$). Scale bar, 100 μ m. NS, $P=0.500$, Student's two-sided t -test. Data are mean \pm s.d. **c**, Flow cytometry analysis showing the efficiency of clodronate in depleting Kupffer cells ($CD45^+F4/80^+$) after lipopolysaccharide (LPS) treatment ($n=3$). Bottom, representative micrographs of HNF4 α and K19 immunostaining analysis of Epo-induced

tumours with and without Kupffer cell depletion ($n=3$). Scale bar, 100 μ m. **d**, Quantifications of liver-infiltrating immune cells from Fig. 4a, b. B220 $P=0.6255$, CD3 $P=0.7649$, Ly6G $P=0.3966$, MHCII $P=0.9889$, Student's two-sided t -test. Data are mean \pm s.d. **e**, Quantification of T cells ($CD45^+CD3^+$, $P=0.2622$), T-helper cells ($CD45^+CD3^+CD8^-CD4^+$, $P=0.960$) and killer T cells ($CD45^+CD3^+CD8^+CD4^-$, $P=0.0914$) ($n=6$). P values determined by Student's two-sided t -test. Data are mean \pm s.d. **f**, Quantification of monocytic immature myeloid cells (moIMC; $CD11b^+Gr1^{low}Ly6c^+F4/80^-$, $P=0.0750$), neutrophilic immature myeloid cells (NeuIMC; $CD11b^+Gr1^+Ly6c^-F4/80^-$, $P=0.2483$) and macrophages ($CD11b^+Gr1^-Ly6c^-F4/80^+$, $P=0.1744$) ($n=3$). P values determined by Student's two-sided t -test. Data are mean \pm s.d.



Extended Data Fig. 5 | Induction of hepatocyte cell death after HDTV or Epo. **a**, Representative micrographs of TUNEL (red) and DAPI (blue) staining in livers of $ROSA^{mT/mG} \times Alb-cre \times p19^{Arf-/-}$ mice with native membrane GFP (green) in hepatocytes 3 days after Epo or HDTV transfection ($n = 3$). Scale bars, 100 μ m. **b**, *Ripk3* mRNA expression in

$p19^{Arf-/-}$ livers 3 days after HDTV delivery of pCaMIN compared to Epo delivery of pCaMIN, determined by qRT-PCR ($n = 4$). $*P = 0.0485$, Student's two-sided t -test. Data are mean \pm s.d. **c**, Representative immunohistochemistry of RIPK3 in livers 3 days after Epo or HDTV treatment ($n = 3$). Scale bars, 100 μ m.

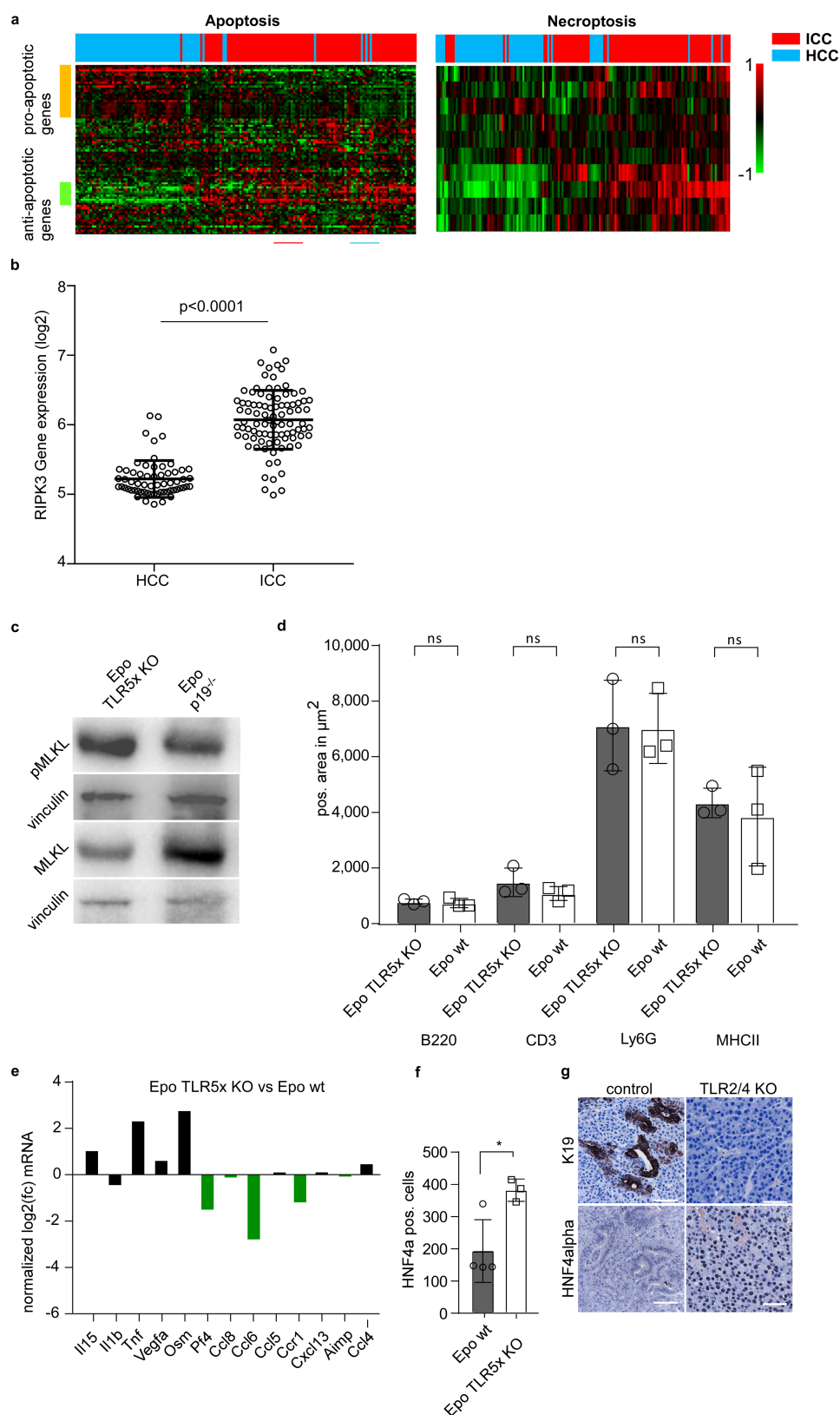


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Necroptotic cell death affects the hepatic microenvironment and tumorigenesis.

a, Representative TUNEL (green) and DAPI (blue) staining in liver sections from mice with ($n = 4$) or without ($n = 3$) Nec-1 pre-treatment 3 days after Epo transfection. Scale bar, 100 μm . **b**, Quantification of TUNEL-positive cells from mice with ($n = 4$) or without ($n = 3$) Nec-1 pre-treatment 3 days after Epo transfection. * $P = 0.0264$, Student's two-sided t -test. Data are mean \pm s.d. **c**, Western blot analysis for the apoptosis marker cleaved caspase 3 in liver lysates from livers with ($n = 4$) or without ($n = 3$) Nec-1 pre-treatment 3 days after Epo transfection. **d**, Western blot analysis for MLKL and pMLKL in liver lysates from livers with ($n = 4$) or without ($n = 3$) Nec-1 pre-treatment 3 days after Epo transfection. **e**, Immunohistochemistry quantification of B220 ($P = 0.7745$), CD3 ($P = 0.9809$), Ly6G ($P = 0.0075$) or MHCII ($P = 0.0994$) in livers with or without Nec-1 pre-treatment 3 days after Epo transfection ($n = 3$). P values determined by Student's two-sided t -test. Data are mean \pm s.d. **f**, Magnification of photographs depicted in Fig. 4k, right. Quantification of HNF4 α -positive cells in Epo-induced tumours with or without Nec-1 pre-treatment ($n = 4$). * $P = 0.0407$, Student's two-sided t -test. Data are mean \pm s.d. **g**, Western blot analysis of MLKL on lysates from hepatocytes isolated via perfusion from $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ or $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice. The experiment was done once with two independent $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice and one $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ mouse. **h**, Western blot analyses for MLKL, pMLKL and vinculin on lysates from $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice 3 days after Epo treatment. Depicted blot is as shown in Fig. 4d (bottom), with an additional lane showing the pMLKL signal obtained in $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ mice 3 days after Epo treatment. The experiment was performed twice with similar results. **i**, Quantification of the duration until tumour size exceeds 0.5 cm after Epo delivery of pCaMIN in

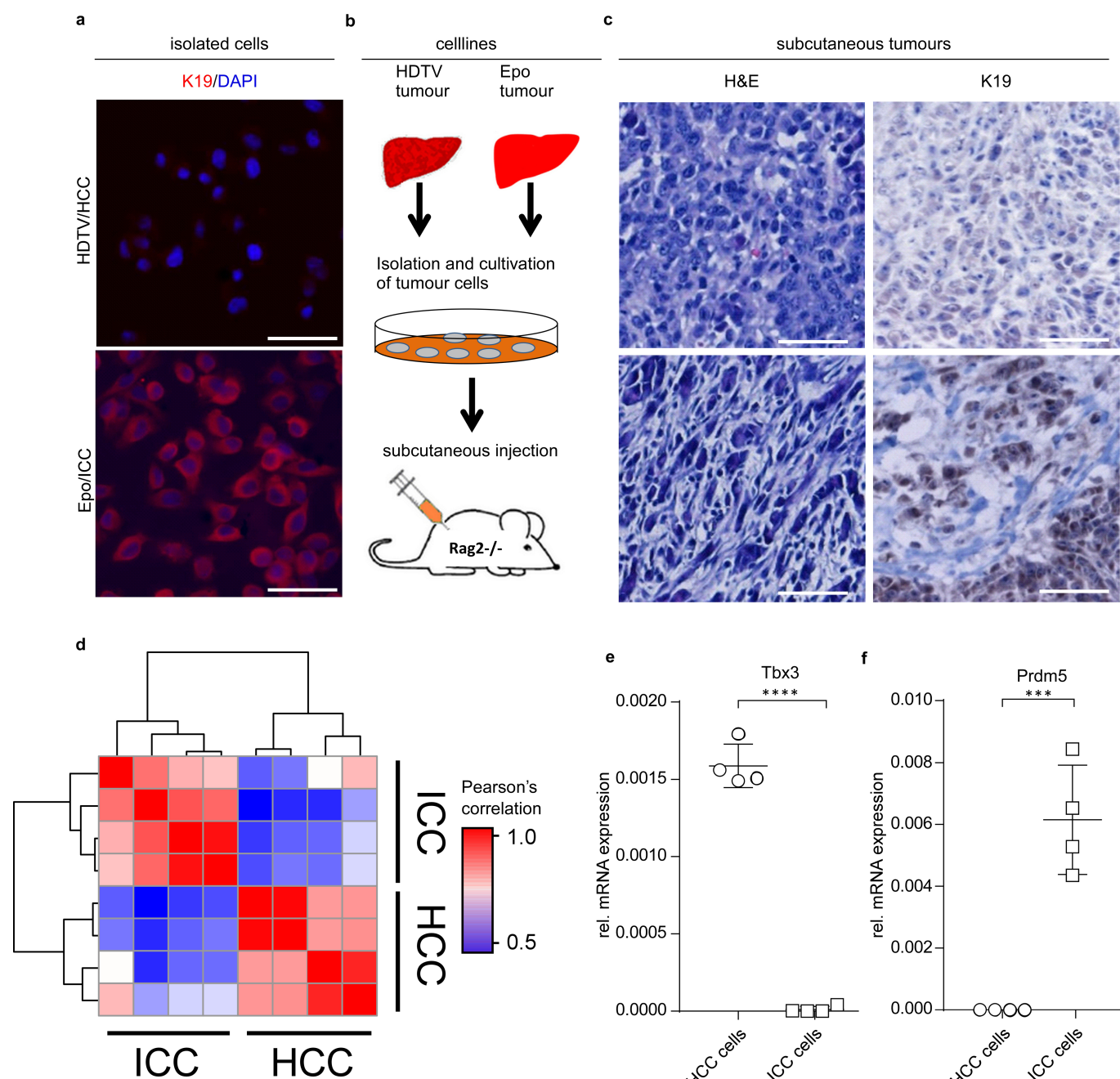
$p19^{Arf^{-/-}}$ mice or pCaMIN plus Cas9n and sgRNA against $p19^{Arf}$ in wild-type mice ($n = 7$). NS, $P = 0.0913$, Student's two-sided t -test. Data are mean \pm s.d. **j**, Immunohistochemistry quantification of B220 ($P = 0.9220$), CD3 ($P = 0.1577$), Ly6G ($P = 0.2375$) or MHCII ($P = 0.3870$) in liver sections from $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ or $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice 3 days after Epo treatment ($n = 3$). P values determined by Student's two-sided t -test. Data are mean \pm s.d. **k**, qPCR-based necroptosis-associated cytokine profile measured on mRNA isolated from livers of $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ or $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice 3 days after Epo treatment. Overlapping downregulated cytokines with Nec-1-treated mice are indicated in green (compare to Fig. 4g). From the 11 cytokines that were found to be suppressible by Nec-1 treatment (Fig. 4g), the expression of 6 was found to be attenuated in Epo-treated MLKL-deficient livers as compared to wild-type livers. This difference might be explained by Nec-1-mediated inhibition of RIPK1-dependent signalling in cells other than hepatocytes. This could also explain why Nec-1 treatment reduced the Ly6G-positive cells in Epo livers (compare to Extended Data Fig. 6e), whereas MLKL deficiency had no effect on the numbers of Ly6G-positive cells after Epo treatment (compare to Extended Data Fig. 6j) ($n = 2$). Data are fold change of the mean from each group. **l**, Quantification of HNF4 α -positive cells in liver sections of Epo-induced tumours in $Mkl^{fl/fl} \times Alb\text{-}cre^{-/-}$ or $Mkl^{fl/fl} \times Alb\text{-}cre^{+/+}$ mice ($n = 5$). * $P = 0.0381$, Student's two-sided t -test. Data are mean \pm s.d. **m**, Representative photograph of HNF4 α and K19 staining of pCaMIN Epo-derived tumours in Mkl wild-type \times $Alb\text{-}cre^{+/+}$ mice ($n = 2$). **n**, Representative micrographs of pRIPK3 immunohistochemistry in tissue sections from sham-operated or bile duct ligated livers of $Arp^{19^{-/-}}$ mice ($n = 3$ each). Scale bars, 100 μm . **o**, Western blot analyses for MLKL and pMLKL on liver lysates from sham-operated or bile duct ligated $Arp^{19^{-/-}}$ mice ($n = 3$ each).



Extended Data Fig. 7 | See next page for caption.

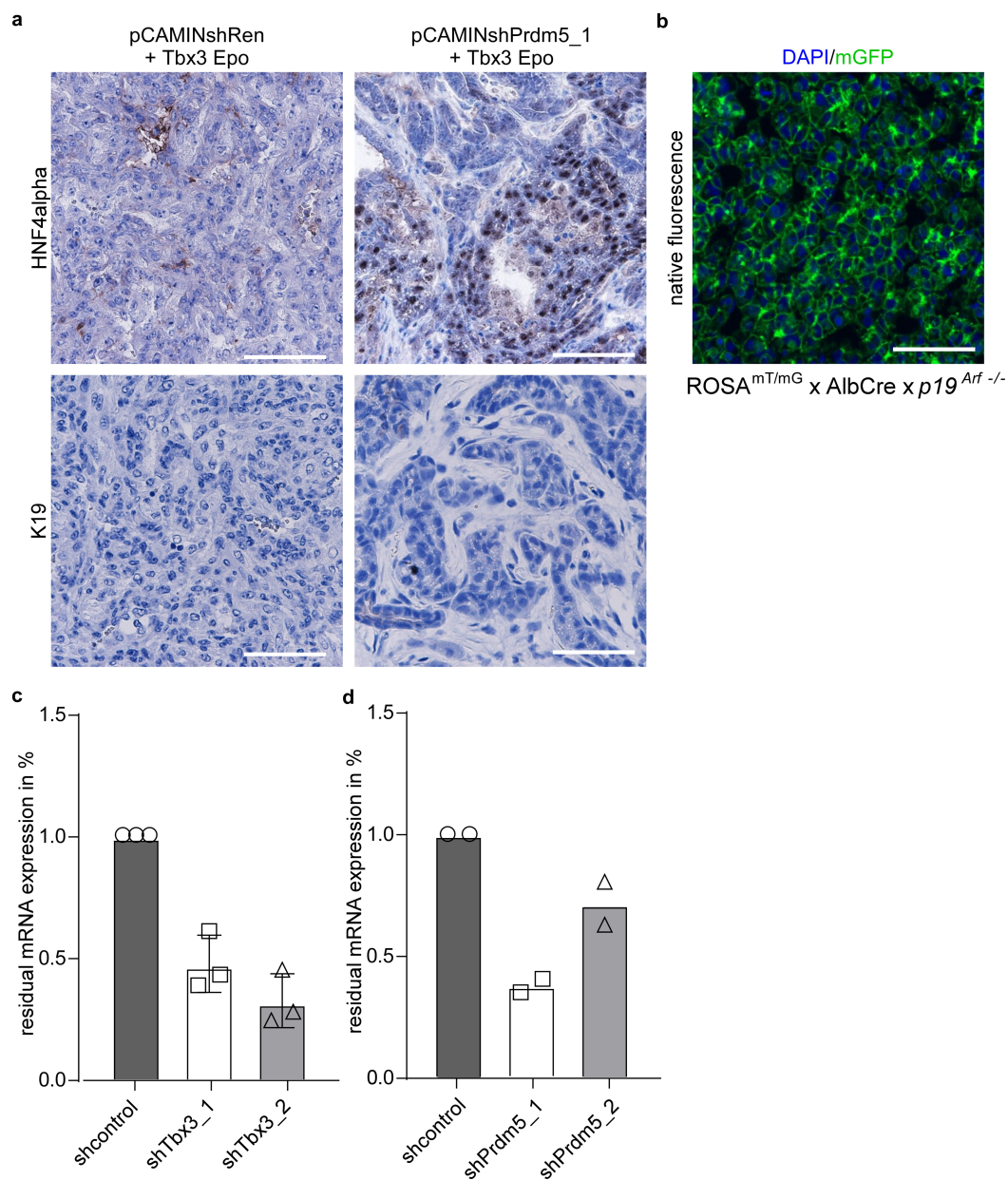
Extended Data Fig. 7 | Necroptosis signatures are found in primary human liver carcinomas. **a**, Transcriptomic patterns of apoptosis- ($n = 84$) or necroptosis- ($n = 10$) related genes in patients with HCC and ICC ($n = 199$) analysed via hierarchical clustering analysis. **b**, Gene expression of *RIPK3* in ICC and HCC patient samples from the TIGER-LC cohort²⁴ ($n = 199$). $P < 0.0001$, Student's two-sided t -test. Data are mean \pm s.d. **c**, Western blot analysis for MLKL and pMLKL in lysates from TLR-knockout and *p19^{Arf}*^{-/-} mouse livers 3 days after Epo treatment. The experiment was performed once ($n = 4$ mice each). **d**, Immunohistochemistry quantification of B220 ($P = 0.6698$), CD3 ($P = 0.2846$), Ly6G ($P = 0.9362$) or MHCII ($P = 0.6734$) in livers from

TLR5-knockout or syngeneic wild-type mice 3 days after Epo treatment ($n = 3$). P values determined by Student's two-sided t -test. Data are mean \pm s.d. **e**, qPCR-based cytokine profile of necroptosis-associated pattern in TLR5-knockout or syngeneic wild-type mice 3 days after Epo treatment ($n = 2$). Data are fold change of the mean from each group. **f**, Quantification of HNF4 α -positive cells in Epo-induced tumours in TLR KO (TLR2, 3, 4, 7 and 9-knockout) ($n = 3$) or syngeneic wild-type ($n = 4$) mice. * $P = 0.0255$, Student's two-sided t -test. Data are mean \pm s.d. **g**, Representative micrographs of HNF4 α and K19 staining on sections from tumours triggered by pCaMIN Epo delivery in TLR2 and TLR4 knockout or syngeneic wild-type mice ($n = 5$). Scale bar, 100 μ m.



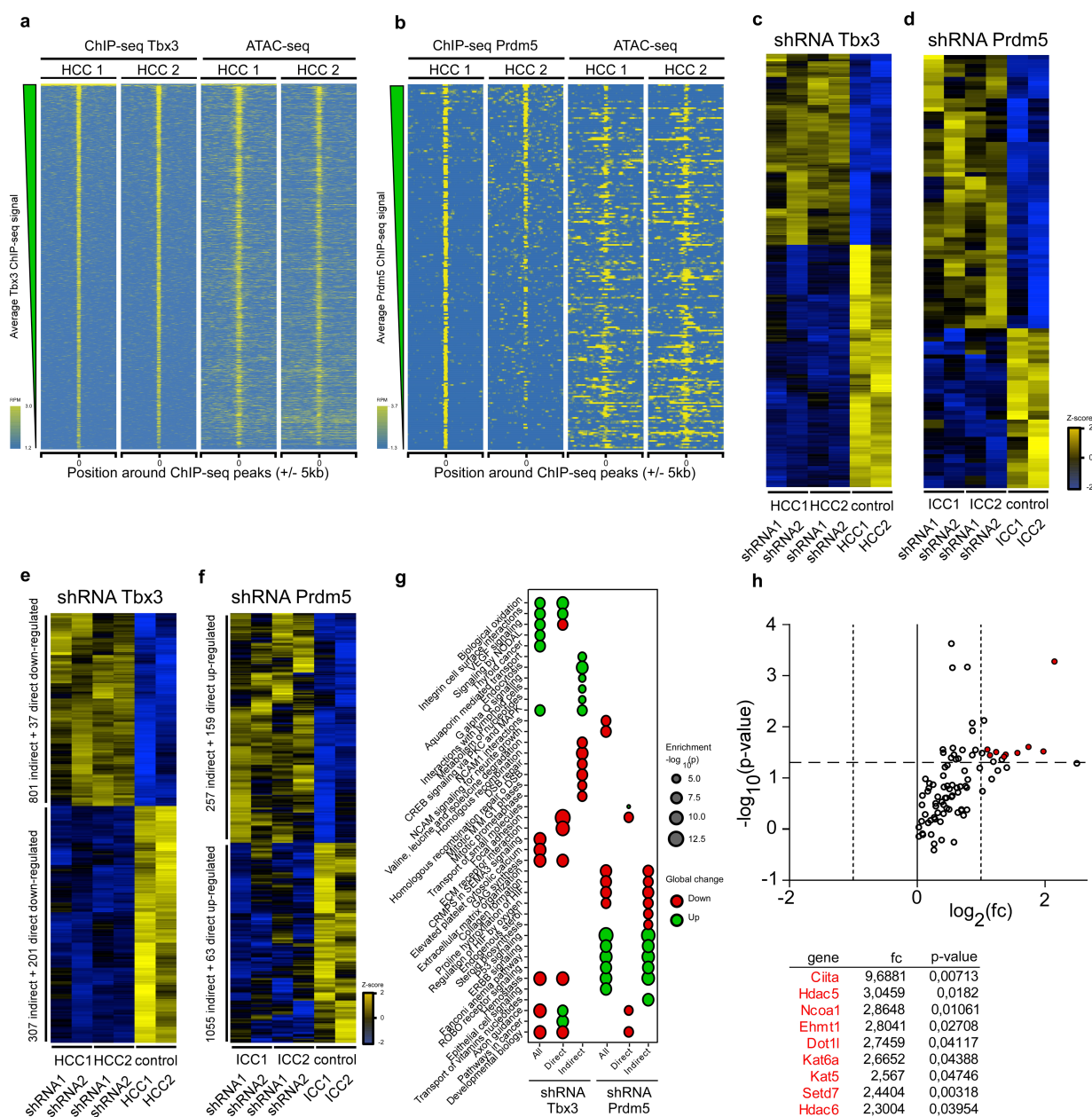
Extended Data Fig. 8 | Generation and analysis of clonally derived cell lines from HDTV or Epo tumours. **a**, Immunocytochemistry of isolated single cell lines of HDTV-derived HCC and Epo-derived ICC tumours. Depicted are representative co-staining images of K19 (red) and DAPI (blue). Scale bars, 100 μ m. Experiment was performed twice with similar results. **b**, Schematic outline of the generation of clonal cell lines of Epo and HDTV tumours for subcutaneous injection into immunodeficient Rag2^{-/-} mice. **c**, Representative micrographs of sections from subcutaneously grown HCC (see **b**; top) and ICC (bottom) with H&E

(left) and K19 (right) staining. These data show that both HCC and ICC phenotypes are stably maintained even after in vitro passaging and in vivo retransplantation procedures in mice ($n = 3$). Scale bars, 100 μ m. **d**, Bi-clustering of pairwise Pearson's correlations based on normalized ATAC-seq fragment pseudo-counts for differentially accessible areas in ICC ($n = 4$ single cell clones) and HCC ($n = 4$ single cell clones). **e**, **f**, qRT-PCR analysis for *Tbx3* (**e**) or *Prdm5* (**f**) in mouse HCC or ICC cells ($n = 4$ single cell clones each). *** $P = 0.0004$, **** $P < 0.0001$, Student's two-sided t -test. Data are mean \pm s.d.



Extended Data Fig. 9 | Influence of PRDM5 and TBX3 on tumour phenotype. **a**, Representative micrographs of immunostaining for HNF4 α or K19 on tumour sections after Epo delivery of pCaMIN transposon vector co-expressing control shRNA (shRen) and full-length *Tbx3* (pCAMINshRen + Tbx3 Epo) or pCaMIN vector co-expressing *Prdm5* shRNA and full-length *Tbx3* (pCAMINPrdm5_1 + Tbx3 Epo) ($n = 3$). Scale bars, 100 μ m. **b**, Representative micrograph of tumours induced by

Epo delivery of pCaMIN and *Tbx3* overexpression in $ROSA^{mT/mG} \times Alb-cre \times p19^{Arf^{-/-}}$ mice showing DAPI (blue) and mGFP (green) positivity ($n = 6$). Scale bar, 100 μ m. **c**, qRT-PCR analysis for *Tbx3* in mouse HCC cells stably expressing shRNAs targeting *Tbx3* (shTbx3_1 and shTbx3_2; $n = 3$). Data are mean \pm s.d. **d**, qRT-PCR analysis for *Prdm5* in mouse ICC cells stably expressing shRNAs targeting *Prdm5* (shPrdm5_1 and shPrdm5_2; $n = 2$). Data are mean \pm s.d.



Extended Data Fig. 10 | Direct and indirect changes of *Tbx3* and *Prdm5* targets and pathways. **a, b**, ChIP-seq density heat map for two biological replicates in the global set of reproducible peaks detected for *Tbx3* (**a**) and *Prdm5* (**b**) following the irreproducible discovery rate workflow (**a** and **b**, left) and corresponding ATAC-seq signal (**a** and **b**, right). Peaks are ranked according to the average ChIP-seq signal across replicates. The data are expressed as normalized reads per million mapped reads (RPM). The signal is shown 5 kb upstream and downstream of the centre of the ChIP-seq peaks. **c, d**, Heat maps depicting gene expression changes after *Tbx3* shRNA-mediated (**c**) and *Prdm5* shRNA-mediated (**d**) suppression. Only direct *Tbx3* and *Prdm5* targets are shown. Data are expressed as *z*-score. For each transcription factor (TBX3 or PRDM5), *n* = 4 cases (2 shRNAs per target, biological duplicates for each) and *n* = 2 controls (1 control shRNA in duplicate), two-sided moderated *t*-statistics. **e, f**, Heat maps depicting gene expression changes after *Tbx3* (**e**) and *Prdm5* (**f**) shRNA-mediated stable knockdown. Each knockdown experiment was performed in established cell lines from two different clones using two different shRNAs. In these heat maps, both direct and indirect *Tbx3* and *Prdm5* ChIP-seq-derived gene targets are shown. Differentially regulated genes were separated into direct or indirect *Tbx3*

or *Prdm5* targets based on the presence or absence of proximal ChIP-seq peaks (<100 kb from the TSS or inside the gene body of deregulated genes). Data are expressed as row *Z*-score. For each transcription factor (TBX3 or PRDM5), *n* = 4 cases (2 shRNAs per target, biological duplicates for each) and *n* = 2 controls (1 control shRNA in duplicate), two-sided moderated *t*-statistics. **g**, Functional over-representation map depicting MSigDB canonical pathways associated to all/direct target/indirect target genes perturbed after *Tbx3* and *Prdm5* knockdown. The size of dots is proportional to the *P* value based on the hypergeometric distribution obtained when testing for over-representation, and their colour denotes whether the term is enriched for up or downregulated gene list. These data show regulation of distinct downstream pathways between *Tbx3* (for example, biological oxidation, developmental biology) and *Prdm5* (for example, extracellular matrix organization, collagen formation or ErbB signalling) (*n* = 4 cases; 2 shRNAs per target, biological duplicates for each, and *n* = 2 controls; 1 control shRNA in duplicate). **h**, qRT-PCR analysis of epigenetic modifiers from livers 3 days after Epo or HDTV treatment. All significantly regulated genes are shown (*n* = 3). *P* values determined by Student's two-sided *t*-test. Data are fold changes of the mean.

The interaction landscape between transcription factors and the nucleosome

Fangjie Zhu¹, Lucas Farnung², Eevi Kaasinen³, Biswajyoti Sahu⁴, Yimeng Yin³, Bei Wei³, Svetlana O. Dodonova², Kazuhiro R. Nitta⁵, Ekaterina Morgunova³, Minna Taipale^{1,3}, Patrick Cramer^{2,5} & Jussi Taipale^{1,3,4*}

Nucleosomes cover most of the genome and are thought to be displaced by transcription factors in regions that direct gene expression. However, the modes of interaction between transcription factors and nucleosomal DNA remain largely unknown. Here we systematically explore interactions between the nucleosome and 220 transcription factors representing diverse structural families. Consistent with earlier observations, we find that the majority of the studied transcription factors have less access to nucleosomal DNA than to free DNA. The motifs recovered from transcription factors bound to nucleosomal and free DNA are generally similar. However, steric hindrance and scaffolding by the nucleosome result in specific positioning and orientation of the motifs. Many transcription factors preferentially bind close to the end of nucleosomal DNA, or to periodic positions on the solvent-exposed side of the DNA. In addition, several transcription factors usually bind to nucleosomal DNA in a particular orientation. Some transcription factors specifically interact with DNA located at the dyad position at which only one DNA gyre is wound, whereas other transcription factors prefer sites spanning two DNA gyres and bind specifically to each of them. Our work reveals notable differences in the binding of transcription factors to free and nucleosomal DNA, and uncovers a diverse interaction landscape between transcription factors and the nucleosome.

The packaging of eukaryotic genomes is accomplished by histones, proteins that form an octameric complex that binds to the DNA backbone, forming nucleosomes^{1–4}. In a canonical nucleosome, a 147-base pair (bp) segment of DNA is wrapped around the histone octamer in a left-handed, superhelical arrangement for a total of 1.65 turns, with the DNA helix entering and exiting the nucleosome from the same side of the histone octamer. The two DNA gyres are parallel to each other except at the position located between the entering and the exiting DNA, where a dyad region of approximately 15 bp contains only a single DNA gyre.

The nucleosome presents a barrier for the binding of proteins such as RNA polymerases to DNA^{5–8}. Similarly, most transcription factors (TFs) are thought to be unable to bind to nucleosomal DNA^{9,10}, except for a specific class of TFs called the pioneer factors¹¹. Despite the importance of the nucleosome in both chromatin organization and transcriptional control^{12–17}, the effect of nucleosomes on the binding of transcription factors has not been systematically characterized.

Nucleosome CAP–SELEX

To determine the effect of nucleosomes on TF–DNA binding, we developed nucleosome consecutive affinity purification–systematic evolution of ligands by exponential enrichment (NCAP–SELEX; Fig. 1a, Extended Data Fig. 1). The method is based on analysis of enrichment of specific sequences from complex 147-bp (lig147) or 200-bp (lig200) DNA libraries, containing 101- or 154-bp randomized regions, respectively. The sequences are reconstituted into a nucleosome, and the complexes incubated with TFs, which are subsequently purified and the bound DNA is recovered using PCR. After multiple selection rounds, the dissociated nucleosomal DNA is separated from intact nucleosomes. Analysis of the sequences enriched by NCAP–SELEX allows inference of TF binding specificities and positions on nucleosomal DNA, together with their effects on the stability of the nucleosome.

We performed SELEX both using nucleosomal (NCAP–SELEX) and free DNA (high-throughput SELEX^{18,19}) using 413 human TF extended DNA binding domains (eDBDs) and 46 full-length constructs (Extended Data Fig. 1h, Supplementary Table 1). The selected TFs covered 29% of the high-confidence TFs from a previously published study²⁰. The enriched sequences were analysed computationally using motif matching, de novo motif discovery and mutual-information (MI) pipelines (see Supplementary Methods). Because nucleosomes can affect TF motifs²¹, we primarily used a MI measure, which can capture any type of enriched sequence pattern (see Fig. 1b). Standard MI analysis also captures nucleosome sequence preference. To separate TF signals from the nucleosome signal, we limited the MI measure to the most highly enriched subsequences (enriched-sequence-based MI; E-MI; Fig. 1b). In parallel, we also analysed all data using motif-based approaches to explain and validate the findings (Supplementary Data 1, 2). Among the tested TFs, 220 eDBDs and 13 full-length constructs were successful (Fig. 1c; see Supplementary Methods for details).

Nucleosome inhibits TF binding

To determine the general effect of nucleosomes on TF–DNA binding, we analysed E-MI signals on lig200, which can accommodate only one nucleosome and contains both nucleosomal and free DNA (Fig. 2a, Extended Data Figs. 2, 3). On lig200 almost all TFs had a lower E-MI signal at the centre (Extended Data Fig. 2a), where the nucleosome occupancy is highest, indicating that the DNA-binding of most TFs is inhibited or spatially restricted by the presence of a nucleosome. However, the effect of the nucleosome on TF binding varied strongly between the TFs (Extended Data Fig. 2b, c). For example, SREBF2, RFX3 and JUND2 only show E-MI signal at the extreme ends of the ligand, suggesting that in the presence of free DNA, they are largely excluded from nucleosomal DNA. By contrast, other TFs

¹Department of Biochemistry, University of Cambridge, Cambridge, UK. ²Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Göttingen, Germany. ³Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. ⁴Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland. ⁵Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. *e-mail: ajt208@cam.ac.uk

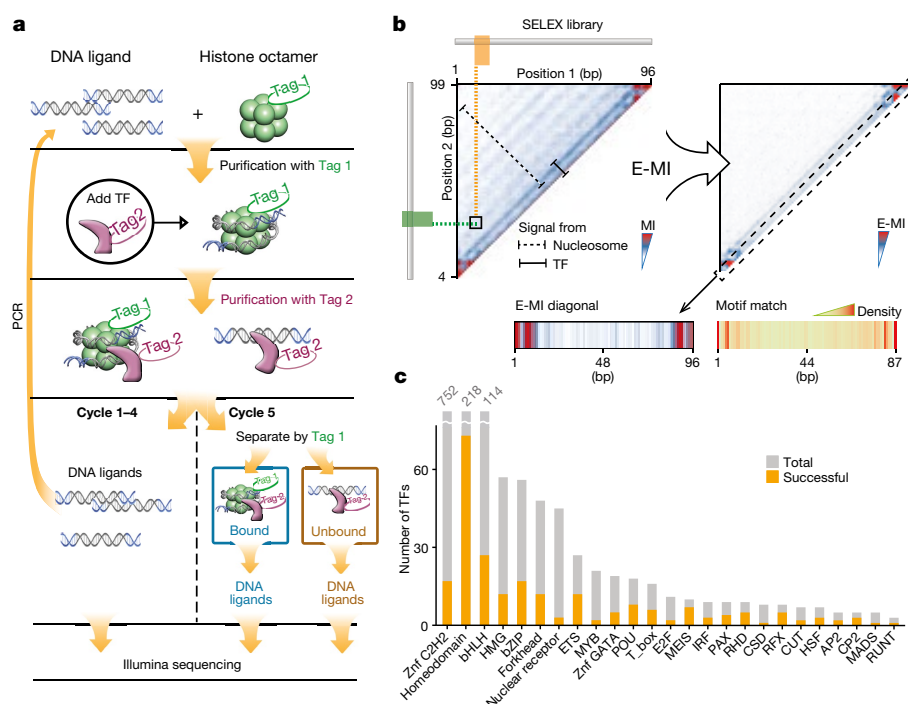


Fig. 1 | Nucleosome CAP-SELEX. a, Schematic representation of NCAP-SELEX. The DNA ligands for SELEX contain a randomized region (grey) with fixed adaptors (blue). The protocol first selects ligands that are favoured by the nucleosome, and then from the nucleosome-bound ligand pool selects ligands that bind to a given TF. The orthogonal tagging of histone H2A (tag 1) and TFs (tag 2) enables the consecutive affinity purification. In the last (5th) cycle, the TF-bound DNA ligands are further separated into nucleosome-bound and unbound libraries before sequencing. **b**, TF-signal analysis by E-MI. Both the TF-binding

signals (solid bar) and the nucleosome-binding signals (dotted bar) can be captured by the MI between 3-mer distributions at two non-overlapping positions of the ligand (left). In our analysis, we further focus on MI of the most enriched 3-mer pairs (E-MI, right) to filter out the nucleosome signals. Most analyses in this manuscript use the E-MI diagonal (box, containing E-MI from directly adjacent non-overlapping 3-mer pairs) because it contains the majority of TF binding signals and is generally similar to the motif-matching result (bottom). **c**, Family-wise coverage of successful TFs.

such as VSX1, ARX, EN1 and SOXs are more capable of binding to nucleosomal DNA. The biochemical ability of TFs to bind to nucleosomal DNA affected their binding also *in vivo* in K562 cells (Extended Data Fig. 2d). These results indicate that the nucleosome often inhibits TF-DNA binding, but that the extent of the effect varies greatly between TFs.

TFs can bind both nucleosomal DNA gyres

Some chromatin-modifying enzymes²² and synthetic molecules²³ can bind both DNA gyres wrapped around the nucleosome. To explore whether TFs can also exhibit such a binding mode, we analysed the entire 2D E-MI signals. We found that binding of the T-box family TF brachyury (T) to nucleosomal DNA resulted in two prominent E-MI signals (Fig. 2b). One was located at the E-MI diagonal (that is, it was observed between adjacent subsequences), whereas the other resulted from sequences located approximately 80 bp from each other. The first signal represents binding of T to nucleosomal DNA similarly to free DNA. The second is associated with an approximately 80-bp motif, indicating dimeric binding that spans both DNA gyres (Fig. 2c). This type of binding was also observed for lig147 but not detected on free DNA (Extended Data Fig. 2e). The signal for the long motif is stronger on the ligands that remained bound to the nucleosome (Extended Data Fig. 2f), indicating that the gyre-spanning mode of T stabilizes nucleosomes. Similar binding was also observed for another T-box factor, TBX2 (Extended Data Fig. 2g), but not for other TFs. Despite the clear biochemical ability of T and TBX2 to bind to nucleosomal DNA using the cross-gyre motif, we did not identify this motif from available ChIP-seq data^{24,25}. Thus, the biological role, if any, of this binding mode needs to be addressed by further experimentation. For some TFs, we also identified weak signals for another binding mode, in which the TFs contact nucleosomal DNA at positions spaced approximately 40-bp apart (for example,

TBX2 and ETV; Extended Data Fig. 2g). These results indicate that the nucleosome scaffold enables new binding modes for TFs that are not possible on free DNA.

Nucleosome affects the orientation of TF binding

In analysis of motif matches on lig200, we noted that the motifs of some TFs displayed a bias of matches in one orientation at the 5' end, and in the other orientation at the 3' end of the ligand. This pattern was observed for many ETS and CREB bZIP factors (Fig. 2d, e, Extended Data Fig. 4). The orientational preference induced by the nucleosome can be explained by the fact that nucleosome breaks the rotational symmetry of DNA (Extended Data Fig. 4d). Depending on TF orientation, a particular side of a TF will be in proximity with either the second gyre of nucleosomal DNA or the histone proteins.

To determine whether the directional binding of TFs to a nucleosome is also observed *in vivo*, we mapped nucleosome positions genome-wide in the human colorectal cancer cell line LoVo using micrococcal nuclease digestion with sequencing (MNase-seq). We found that the nucleosome distribution is asymmetric ($P < 0.0003$, two-sided *t*-test) around ELF1 and ELF2 *in vivo* sites (Fig. 2f, Extended Data Fig. 4e). Such asymmetry is not observed for the same ELF2 sites after salt treatment that laterally mobilizes the nucleosomes, or around ELF2 motif matches that do not show ChIP-seq signal (Fig. 2f). The nucleosome occupancy is lower upstream than downstream of the ELF2 sites. This pattern suggests that the more stable binding of ELF2 downstream of the nucleosome displaces the nucleosome or pushes it upstream. Several chromatin features that are asymmetric relative to sites occupied by TFs have been reported^{26–28}. Our observation that nucleosome itself induces asymmetry in the preferred binding orientation of TFs provides a potential mechanistic basis for these findings.

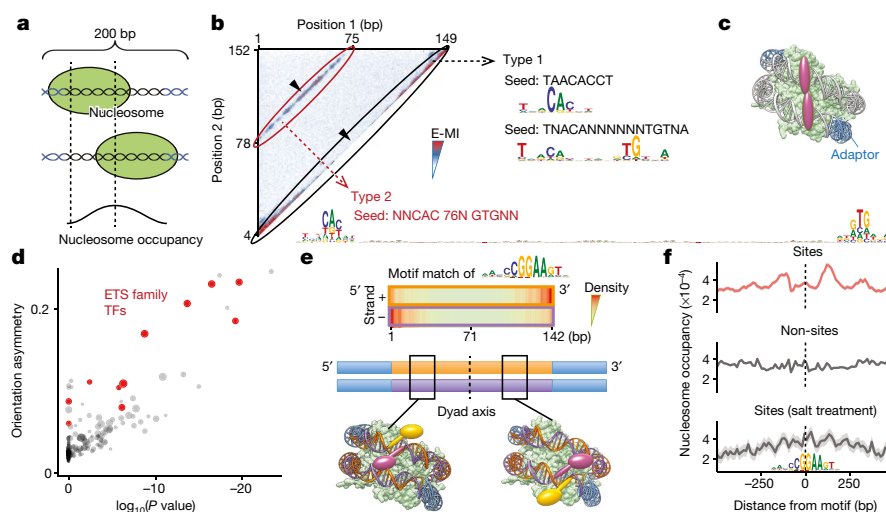


Fig. 2 | Nucleosome scaffolds DNA and breaks its rotational symmetry, enabling new TF binding modes. **a**, Schematic of single nucleosomes assembled on different positions of lig200, resulting in higher nucleosome occupancy towards the centre. **b**, Two different binding types of T (brachyury) on nucleosomal DNA. Heat map shows E-MI for all combinations of positions on lig200. Type 1 signal near the diagonal yields short motifs similar to those on free DNA. The type 2 signal corresponds to a motif approximately 80-bp long. Note that in contrast to type 1 signal, type 2 signal is not inhibited by the high nucleosome occupancy at the centre (arrowheads). **c**, Schematic of TFs (purple) that bind both gyres of nucleosomal DNA. **d**, Orientational asymmetry of binding of individual TFs on nucleosomal DNA. *y* axis: binding energy difference between two relative orientations of the most enriched subsequences. *x* axis: *t*-test *P* value of the difference compared to binding on free DNA (see Supplementary Methods). Note that most ETS-family TFs (red) show prominent asymmetry. Dot size represents the extent of signal enrichment in the NCAP-SELEX library of each TF. **e**, Orientational

asymmetry of the ETS factor ELF2. At the 5' end of the ligand, the ELF2 motif (top) is enriched on the minus strand, because ELF2 prefers to bind DNA in one orientation relative to the nucleosome (yellow, bottom left cartoon). At the 3' end of the ligand, the ELF2 motif is enriched on the plus strand, as this leads to the same orientation of the ELF2 protein with respect to the nucleosome (yellow, bottom right). Note also that the two yellow ELF2 proteins make symmetric contacts, but to different strands of DNA (marked orange and purple; adaptors are indicated in blue). Note that TF positions on the ligand are not fixed, for simplicity only a few example positions are shown. **f**, Asymmetric nucleosome distribution around genomic ELF2 sites (top, sites positioned at centre). Asymmetry is not observed for the same ELF2 sites after salt treatment to mobilize the nucleosome (bottom) or for ELF2 motifs without ChIP signal (middle). Nucleosome positions are shown as frequency of the centre of MNase fragments (140–170 bp). Each profile (*n* = 999 data points) is LOESS smoothed (locally weighted smoothing) with a span of 0.05 and the shaded band indicates the s.e.m.

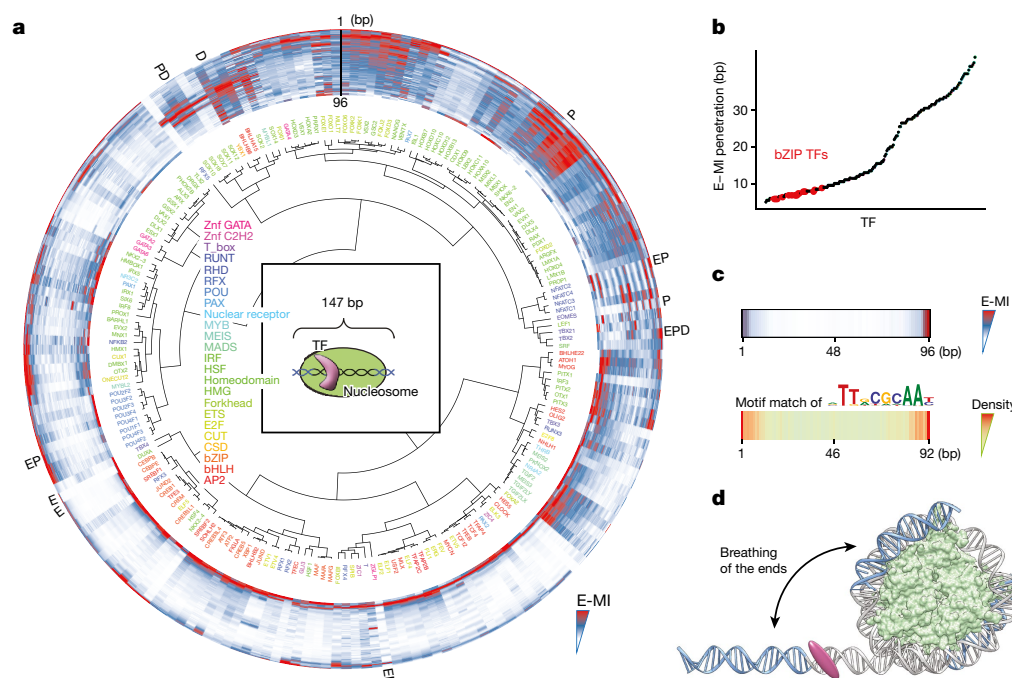


Fig. 3 | Nucleosome induces positional preference to TF binding. **a**, Hierarchical clustering of the E-MI diagonals for NCAP-SELEX with the 147-bp ligand (lig147). E-MI diagonal is scaled for each TF (see Supplementary Methods). The names of the TFs are coloured by family. TFs from the same family tend to be clustered together. A few TFs were annotated as examples to illustrate their end (E), periodic (P) and dyad (D) preferences (see Supplementary Table 5). Note that the

preferences are not mutually exclusive. Centre, schematic of the fixed position of nucleosome on lig147. **b**, E-MI penetration of each TF on lig147. All bZIP TFs are marked with red. **c**, E-MI diagonal and motif-matching results for the bZIP factor CEBPB. **d**, Schematic representation showing a TF that prefers the ends of nucleosomal DNA owing to breathing. Both ends of nucleosomal DNA will breathe but only one is illustrated here for clarity.

Nucleosome induces positional TF-binding preferences

Next we analysed the positional preference of TF binding to nucleosomal DNA. We designed the 147-bp NCAP-SELEX ligand (lig147) that matches the preferred length of nucleosomal DNA²⁹, allowing more precise mapping of TF-binding positions relative to the nucleosome. The results indicate that the presence of nucleosome restricts TF binding, and induces several types of positional preference (Fig. 3, Extended Data Figs. 5, 6). Expert analyses and machine learning analyses (Extended Data Fig. 6b, c, Supplementary Methods) revealed three types of positional preference on nucleosomal DNA (Fig. 3a, Supplementary Table 5): 1) end preference, these TFs prefer positions towards the end of the ligand that are partially accessible due to a process known as ‘breathing’^{1,30,31}. Many TFs of this class either radially cover more than 180° of the DNA circumference (for example, bZIP and bHLH), and/or bind to long motifs through a continuous interaction with DNA (for example, C2H2 zinc fingers) (Fig. 3a); 2) periodic preference, these TFs tend to bind to periodic positions on nucleosomal DNA and 3) dyad preference, these TFs prefer to bind to nucleosomal DNA near the dyad position.

Half of the circumference of nucleosomal DNA is in close proximity to the histones. As DNA is helical, equivalent positions that could be accessible to TFs are located at approximately 10-bp intervals. Accordingly, we found that many TFs prefer to bind to positions located approximately 10 bp apart on nucleosomal DNA (Fig. 3a, Extended Data Fig. 7). By applying a fast Fourier transform (FFT) to the E-MI diagonals, we obtained both the strength and rotational position (phase) of the approximately 10-bp periodicity for each TF (Fig. 4a). Analysis of the rotational position of binding for the TFs revealed that both major and minor grooves of nucleosomal DNA were accessible from the solvent side. For example, PITX and EOMES prefer almost opposite phases (Fig. 4a). This is consistent with the known structures; PITX contacts DNA principally via the major groove³² (structure in Fig. 4b), whereas T-box TFs such as EOMES contact DNA mainly via the minor groove^{33,34} (Extended Data Fig. 7b). Such periodic preference of binding has been reported previously for p53 and the glucocorticoid receptor^{35,36}, but the prevalence of this phenomenon was unclear. Among the TF families, periodic binding was particularly common among homeodomain TFs (Fig. 3a), and was also detected for homeodomain TFs from mouse liver (Extended Data Fig. 7g). Taken together, the results suggest that consistent with structural data³⁷ (Extended Data Fig. 5a), many TFs can bind nucleosomal DNA from the solvent-accessible side.

Analysis of the positional preference of TFs on nucleosomal DNA also revealed that the dyad region is strongly preferred by some TFs (Fig. 4c–g, Extended Data Fig. 8 and previous work^{38,39}). For example, RFX5 shows very strong binding to the dyad positions of lig147 (Fig. 4c); on the basis of a competition assay, the affinity of RFX5 to dyad positions is higher than to free DNA (Fig. 4c, bottom; Extended Data Fig. 8b). To test whether RFX5 also prefers nucleosomal DNA in vivo, we expressed RFX5 in HEK-293 cells, and then detected nucleosome positions and RFX5-occupied sites using MNase-seq and MNase-ChIP. HEK-293 cells do not endogenously express RFX5, and in untransfected cells the positions at which exogenous RFX5 binds are located at a maximum of nucleosome occupancy (Fig. 4d, Extended Data Fig. 8). However, upon RFX5 expression, RFX5 forms a complex with nucleosomes, in which the positions of the nucleosomes are shifted to the sides of the sites that are bound by RFX5 (Fig. 4d, e). These results indicate that RFX5 prefers nucleosomal DNA in vivo, and that it potentially can induce nucleosome remodelling. In addition to RFX5, we also found that multiple SOX TFs have a preference for binding to dyad DNA (Fig. 4f, g). Such a preference was validated for SOX11 using an electrophoretic mobility shift assay (EMSA; Extended Data Fig. 8). Taken together, our results indicate that on nucleosomal DNA, some TFs display a strong preference towards the dyad region.

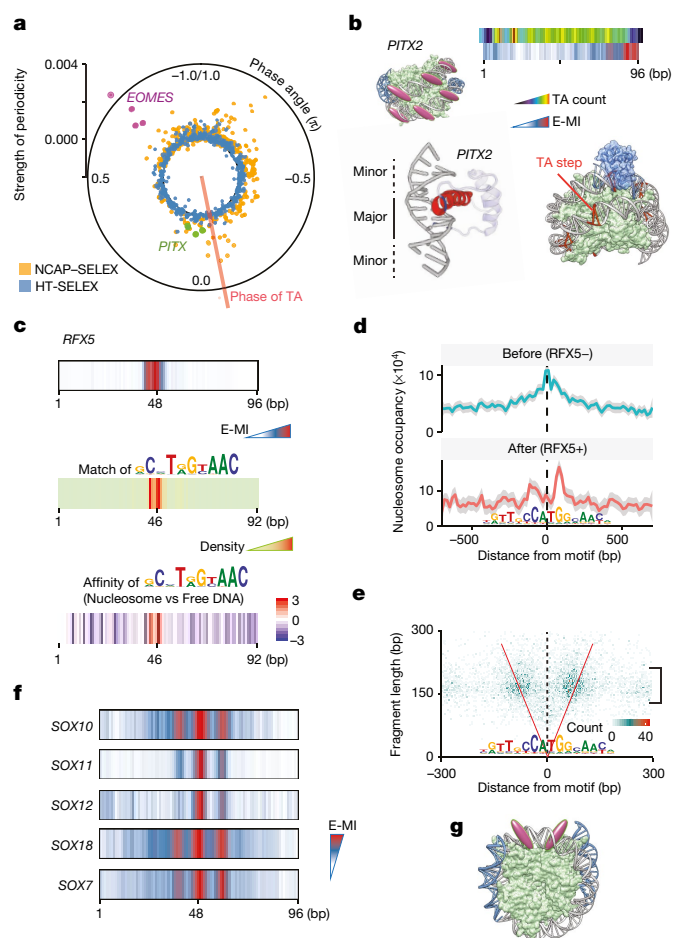


Fig. 4 | Periodic and position-specific binding of TFs to nucleosomal DNA. **a**, TF binding on nucleosomal DNA commonly displays a 10-bp periodic pattern. The polar plot shows strength and phase derived from FFT of E-MI diagonals, for both NCAP-SELEX (orange) and high-throughput SELEX (blue; free DNA). Note that EOMES (magenta, four replicates) and PITX (green for PITX1, 2, 3) have opposite phases. Phase of TA dinucleotide (red line) indicates where the major groove faces outward⁴⁰. **b**, PITX prefers exposed major grooves on nucleosomal DNA. The E-MI diagonal of PITX is in phase with the TA peaks along the ligand. Accordingly, the structure of PITX (PDB entry 2LKX) shows contacts with DNA principally in the major groove. The base-contacting helices (red) and loops (dark blue) are indicated. Cartoon representation to the right shows that the steric hindrance is minimal when PITX (blue) binds in phase with TA (orange) on the nucleosome structure (PDB entry 3UT9). **c**, RFX5 prefers to bind near the nucleosome dyad. E-MI diagonal (top), motif matching (middle), and competition assay (bottom) are shown. Positive values in the competition assay indicate preference towards nucleosomal DNA. **d**, Binding of RFX5 affects local nucleosome profile in vivo. Nucleosome distribution is examined by MNase-seq before (top) and after (bottom) exogenous expression of RFX5 in HEK293 cells. RFX5 motif matches within MNase-ChIP peaks are centred. Nucleosome occupancy is shown as frequency of the centre of MNase fragments (140–170 bp). Each profile ($n = 1,401$ data points) is LOESS smoothed with a span of 0.05 and the shaded band indicates s.e.m. Before RFX5 expression, the nucleosome occupancy is higher at the RFX5 sites than the surrounding region (top); the nucleosomes are shifted after the expression of RFX5 (bottom). **e**, MNase-ChIP indicates that RFX5 binds to nucleosomal DNA in vivo. Counts of MNase-ChIP fragments are binned to 3 bp by 3 bp bins according to their lengths and centre positions. Note that most immunoprecipitated fragments are approximately 150 bp in size (bracket) and overlap the RFX5 motif (are between the red ‘V’ lines), indicating that RFX5 prefers to bind to nucleosomal DNA. **f**, E-MI diagonal of SOX family TFs showing preferred binding around the dyad. **g**, Schematic of TFs that prefer to bind around the dyad.

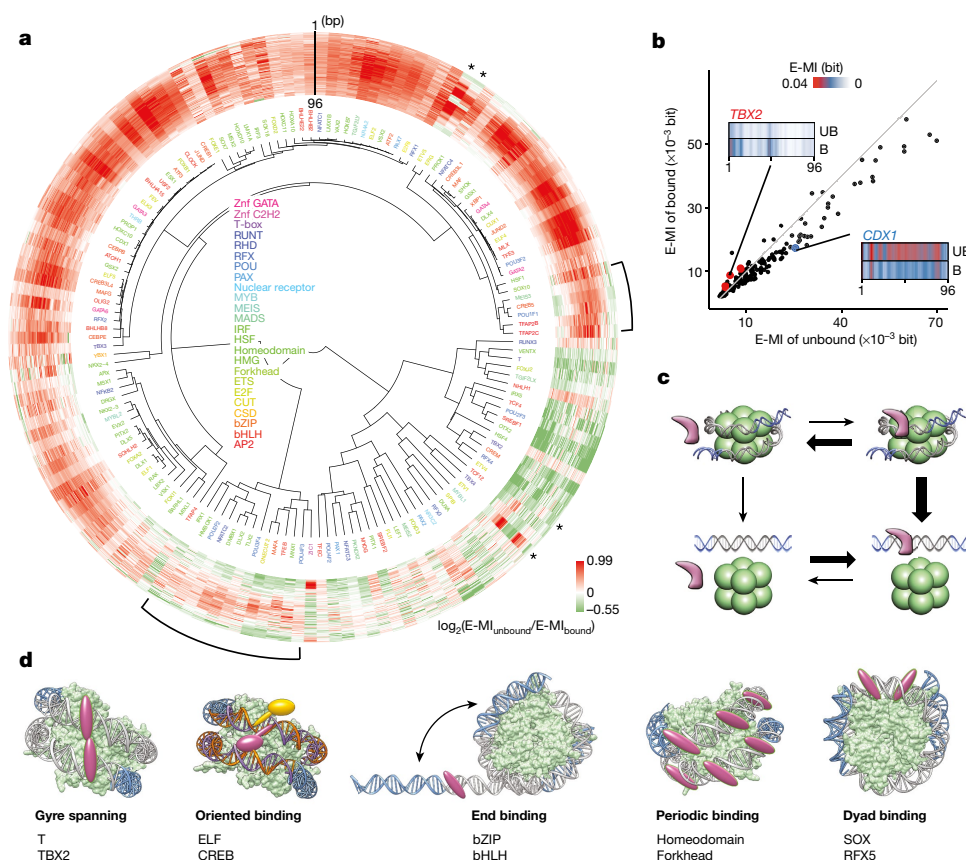


Fig. 5 | Effects of TF binding on nucleosome stability. **a**, Hierarchical clustering of the differential E-MI diagonal between the nucleosome-bound and unbound cycle 5 libraries. Most TFs have stronger signal in the unbound library, indicating that their binding destabilizes the nucleosome. Brackets denote TFs that both destabilize and stabilize the nucleosome in a position-dependent way. Asterisks denote the ETS factors with a specific pattern of positional dependence. **b**, Mean strengths of E-MI diagonals in the nucleosome-bound and unbound cycle 5 libraries. The scatter plot shows the mean E-MI for the diagonals of each TF (dots), and for both the bound library (y axis) and the unbound library (x axis). The grey line represents where $y = x$. Most TFs have stronger signals

Effect of TF binding on nucleosome dissociation

To determine whether TF binding affects the stability of the nucleosome, we performed an additional affinity capture step to separate the nucleosome-bound and dissociated DNA (unbound) in the last cycle of lig147 NCAP-SELEX (Figs. 1a, 5, Extended Data Fig. 9). Control experiments lacking TFs showed very little difference between the E-MI signal of the bound and unbound libraries, whereas in the presence of TFs, clear differences were observed (Fig. 5a, Extended Data Fig. 9a). We found that most TFs (for example, CDX1) have stronger E-MI in the unbound library compared to that of the bound library, suggesting that they can facilitate nucleosome dissociation upon binding (Fig. 5b, c). However, we also identified a few exceptional TFs, the binding of which stabilized the nucleosome. These include the T-box TFs, such as TBX2 (Fig. 5b). Moreover, the effect of TFs on nucleosome stability is also dependent on their binding mode and position on the nucleosomal DNA (Fig. 5a, Extended Data Fig. 9).

Discussion

TFs and the nucleosome are central elements regulating eukaryotic gene expression. In this study, we developed a new method, NCAP-SELEX, for analysis of nucleosome-TF interactions, and systematically examined the binding preference of 220 TFs on nucleosomal DNA. We identified five major interaction patterns between TFs and the nucleosome (Fig. 5d, Extended Data Fig. 10; Supplementary Table 5). The interaction modes are consistent with structural considerations,

and not mutually exclusive. They include 1) binding that spans the two gyres of nucleosomal DNA; 2) orientational preference; 3) end preference; 4) periodic preference and 5) preferential binding to the dyad region.

Binding of most TFs facilitated the dissociation of nucleosomes. The simplest mechanism to explain this finding is that TFs bind to nucleosomal DNA and form a ternary complex. This complex is relatively unstable because the TFs prefer free DNA over nucleosomal DNA. This difference in affinity provides the free energy that facilitates dissociation of the nucleosome. Although the histone octamer binds 147-bp DNA more strongly than most TFs, within the approximately 10-bp segment that is bound by a TF, the bonds formed by the TF are stronger than those formed by histones. Therefore, binding of a TF to a partially dissociated nucleosome can also prevent rewinding of the TF-bound DNA segment to the nucleosome.

The TFs that facilitate the dissociation of nucleosome function as potential activators that can open chromatin and regulate gene expression. Some TFs, in turn, stabilized the nucleosome. These factors could repress gene expression, or to precisely position nucleosomes at specific genomic loci. Our findings are related to previous analyses that have identified pioneer TFs, which can access nucleosomal DNA¹¹. However, our observations indicate that a binary classification of TFs is not sufficient to capture the complete diversity of the interaction landscape between TFs and the nucleosome. Taken together, our results explain in part the complexity of the relationship between sequence and gene

expression in eukaryotes, and provide a basis for future studies aimed at understanding transcriptional regulation based on biochemical principles.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability

All of the computer programs and scripts used are either published or available upon request.

Data availability

All next-generation sequencing data have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB22684. The relevant processed data are included as Supplementary Information.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0549-5>

Received: 23 September 2017; Accepted: 6 August 2018;

Published online 24 September 2018.

- Andrews, A. J. & Luger, K. Nucleosome structure(s) and stability: variations on a theme. *Annu. Rev. Biophys.* **40**, 99–117 (2011).
- Segal, E. & Widom, J. What controls nucleosome positions? *Trends Genet.* **25**, 335–343 (2009).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- McGinty, R. K. & Tan, S. Nucleosome structure and function. *Chem. Rev.* **115**, 2255–2273 (2015).
- Jin, J. et al. Synergistic action of RNA polymerases in overcoming the nucleosomal barrier. *Nat. Struct. Mol. Biol.* **17**, 745–752 (2010).
- Raveh-Sadka, T. et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44**, 743–750 (2012).
- Teves, S. S., Weber, C. M. & Henikoff, S. Transcribing through the nucleosome. *Trends Biochem. Sci.* **39**, 577–586 (2014).
- Hartog, G. A. Transcription elongation by RNA polymerase II. *Curr. Opin. Genet. Dev.* **13**, 119–126 (2003).
- Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
- Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
- Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr. Opin. Genet. Dev.* **37**, 76–81 (2016).
- Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
- Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl Acad. Sci. USA* **107**, 22534–22539 (2010).
- Boyer, L. A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
- Roy, S. et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
- Stanojevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science* **254**, 1385–1387 (1991).
- Yan, J. et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
- Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
- Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
- Vaquerez, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
- Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
- Nodelman, I. M. et al. Interdomain communication of the Chd1 chromatin remodeler across the DNA gyres of the nucleosome. *Mol. Cell* **65**, 447–459.e6 (2017).
- Edayathumangalam, R. S., Weyermann, P., Gottesfeld, J. M., Dervan, P. B. & Luger, K. Molecular recognition of the nucleosomal “super groove”. *Proc. Natl Acad. Sci. USA* **101**, 6864–6869 (2004).
- Faial, T. et al. Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development* **142**, 2121–2135 (2015).
- Lolas, M., Valenzuela, P. D. T., Tjian, R. & Liu, Z. Charting Brachyury-mediated developmental pathways during early mouse embryogenesis. *Proc. Natl Acad. Sci. USA* **111**, 4478–4483 (2014).
- Kundaje, A. et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* **22**, 1735–1747 (2012).
- Sherwood, R. I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
- Isaac, R. S. et al. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife* **5**, e13450 (2016).
- Poirier, M. G., Bussiek, M., Langowski, J. & Widom, J. Spontaneous access to DNA target sites in folded chromatin fibers. *J. Mol. Biol.* **379**, 772–786 (2008).
- Chaney, B. A., Clark-Baldwin, K., Dave, V., Ma, J. & Rance, M. Solution structure of the K50 class homeodomain PITX2 bound to DNA and implications for mutations that cause Rieger syndrome. *Biochemistry* **44**, 7497–7511 (2005).
- Stirnimann, C. U., Ptschek, D., Grimm, C. & Müller, C. W. Structural basis of TBX5-DNA recognition: the T-box domain in its DNA-bound and -unbound form. *J. Mol. Biol.* **400**, 71–81 (2010).
- Coll, M., Seidman, J. G. & Müller, C. W. Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome. *Structure* **10**, 343–356 (2002).
- Cui, F. & Zhurkin, V. B. Rotational positioning of nucleosomes facilitates selective binding of p53 to response elements associated with cell cycle arrest. *Nucleic Acids Res.* **42**, 836–847 (2014).
- Li, Q. & Wrangé, O. Accessibility of a glucocorticoid response element in a nucleosome depends on its rotational positioning. *Mol. Cell Biol.* **15**, 4375–4384 (1995).
- McGinty, R. K. & Tan, S. Recognition of the nucleosome by chromatin factors and enzymes. *Curr. Opin. Struct. Biol.* **37**, 54–61 (2016).
- Zhou, B. R. et al. Structural mechanisms of nucleosome recognition by linker histones. *Mol. Cell* **59**, 628–638 (2015).
- Iwafuchi-Doi, M. et al. The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Mol. Cell* **62**, 79–91 (2016).
- Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).

Acknowledgements We thank F. Zhong, A. Jolma, J. Zhang and J. Toivonen for valuable suggestions; E. Inns for proofreading; T. Kivioja for critical review of the manuscript and L. Hu, J. Liu and S. Augsten for technical assistance. This work was funded by the EU Horizon 2020 project MRGGrammar (664918), Cancerfonden (120529, 150662), Knut and Alice Wallenberg Foundation (2013.0088), Vetenskapsrådet (D0815201), Academy of Finland CoE (312042) (J.T.); DFG (SFB860, SPP1935), ERC AdG TRANSREGULON (693023), Volkswagen Foundation (P.C.) and EMBO fellowship ALTF 949-2016 (S.D.).

Reviewer information Nature thanks T. Hughes, B. F. Pugh and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.T., F.Z. and P.C. conceived the experiments. F.Z. performed most experiments and analyses. L.F. produced the histone octamers. B.S. and E.K. contributed to generation and analysis of the MNase-seq and ChIP-seq data, respectively. B.W. and S.O.D. performed SOX EMSA and the binding assay with nuclear proteins, respectively. Y.Y. contributed to protein production and motif analysis. M.T., K.R.N. and E.M. contributed to design and analysis of sequencing and structure data. F.Z. and J.T. interpreted the data and wrote the manuscript. All authors discussed the findings and contributed to the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0549-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0549-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.T.

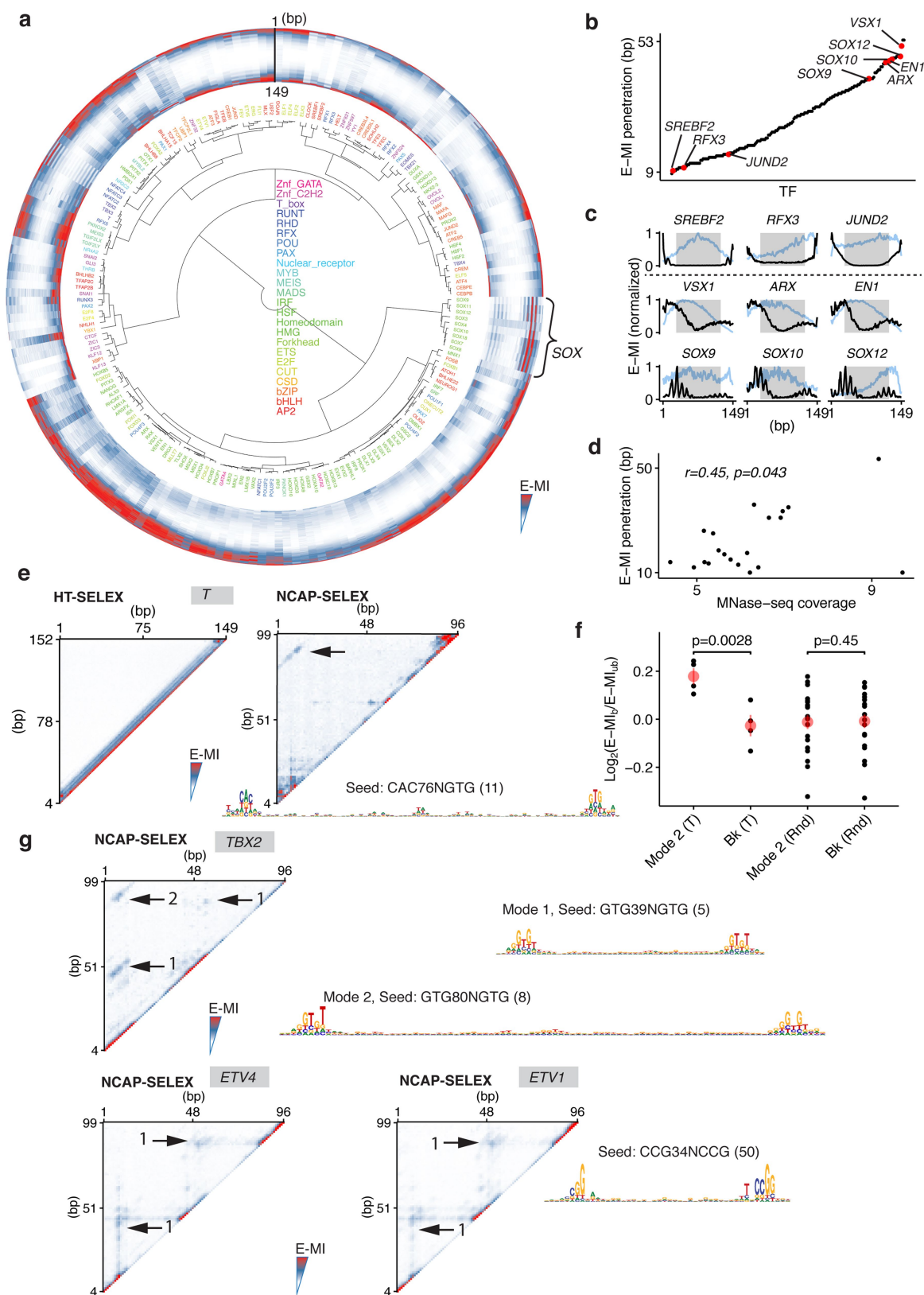
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended Data Fig. 1 | Experiment design and data analysis strategy of NCAP-SELEX.

a, Expression of the recombinant histones from *Xenopus laevis*. For each lane 3 μ g histone is loaded. Similar purifications for untagged H2A, H2B, H3, and H4 have been repeated at least three times. The SBP-H2A purification was performed once. **b**, Size-exclusion chromatogram of the histone octamer. Octamer formation was performed twice and the results were highly consistent. **c**, EMSA result showing the reconstituted nucleosomes using lig147 and lig200. The original ligands are also loaded as reference. The asterisks indicate the nucleosome bands. Similar results are seen in four independent nucleosome reconstitutions. For gel source data see Supplementary Fig. 1. **d**, Oligonucleotide periodicity in the library enriched by nucleosome. As a quality control of nucleosome reconstitution, we verified whether nucleosome by itself is enriching the previously reported approximately 10-bp periodic oligonucleotide signal^{41,42}. Nucleosome SELEX (without TF) were carried out for four cycles to enrich nucleosome-favouring ligands. The counts of each single and di-nucleotide across each individual ligand were Fourier transformed and summed up for the whole library. A clear peak around 0.1 bp⁻¹ (corresponding to the reported approximately 10-bp periodicity) is visible for most mono- and dinucleotides. **e**, The C/G/CG preferences of nucleosome. All 9-mers were counted for the nucleosome-favoured

(bound) and the nucleosome-disfavoured (unbound) libraries. The point representing each 9-mer is coloured according to its C/G/CG content (top), and the count ratios between the bound and the unbound libraries are summarized for 9-mers of different C/G/CG contents (bottom). For the box plots grouped by C/G content, the sample sizes of the boxes are 19,683, 59,049, 78,732, 61,236, 30,618, 10,206, 2,268, 324, 27 and 1, respectively for 9-mer groups containing 0 to 9 C/G. For the box plots grouped by CG dinucleotide content, the sample sizes of the boxes are 151,316, 91,824, 17,784, 1,200 and 20, respectively, for 9-mer groups containing 0 to 4 CG. The line within each box represents the median; the lower and upper boundaries of the box indicate the first and third quartiles and the whiskers represent the 1.5-fold interquartile range. More extreme values are indicated with dots. **f**, Analysis pipeline for the ligands enriched in NCAP-SELEX. **g**, E-MI strength comparison for libraries with and without TF signals. The E-MI heat maps represent signals in the input (cycle 0) library, in the cycle 4 library of nucleosome-favoured sequences (Nucl. SELEX), and in the NCAP- and high-throughput (HT)-SELEX cycle 4 libraries. The libraries enriched with TF (NCAP and HT) have much stronger E-MI signals compared to the cycle 0 and the nucleosome-SELEX library. The detected dimer signals of HSF1 in HT-SELEX is boxed. **h**, Family-wise coverage of TFs tried in NCAP-SELEX.

41. Collings, C. K., Fernandez, A. G., Pitschka, C. G., Hawkins, T. B. & Anderson, J. N. Oligonucleotide sequence motifs as nucleosome positioning signals. *PLoS ONE* **5**, e10933 (2010).
42. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).



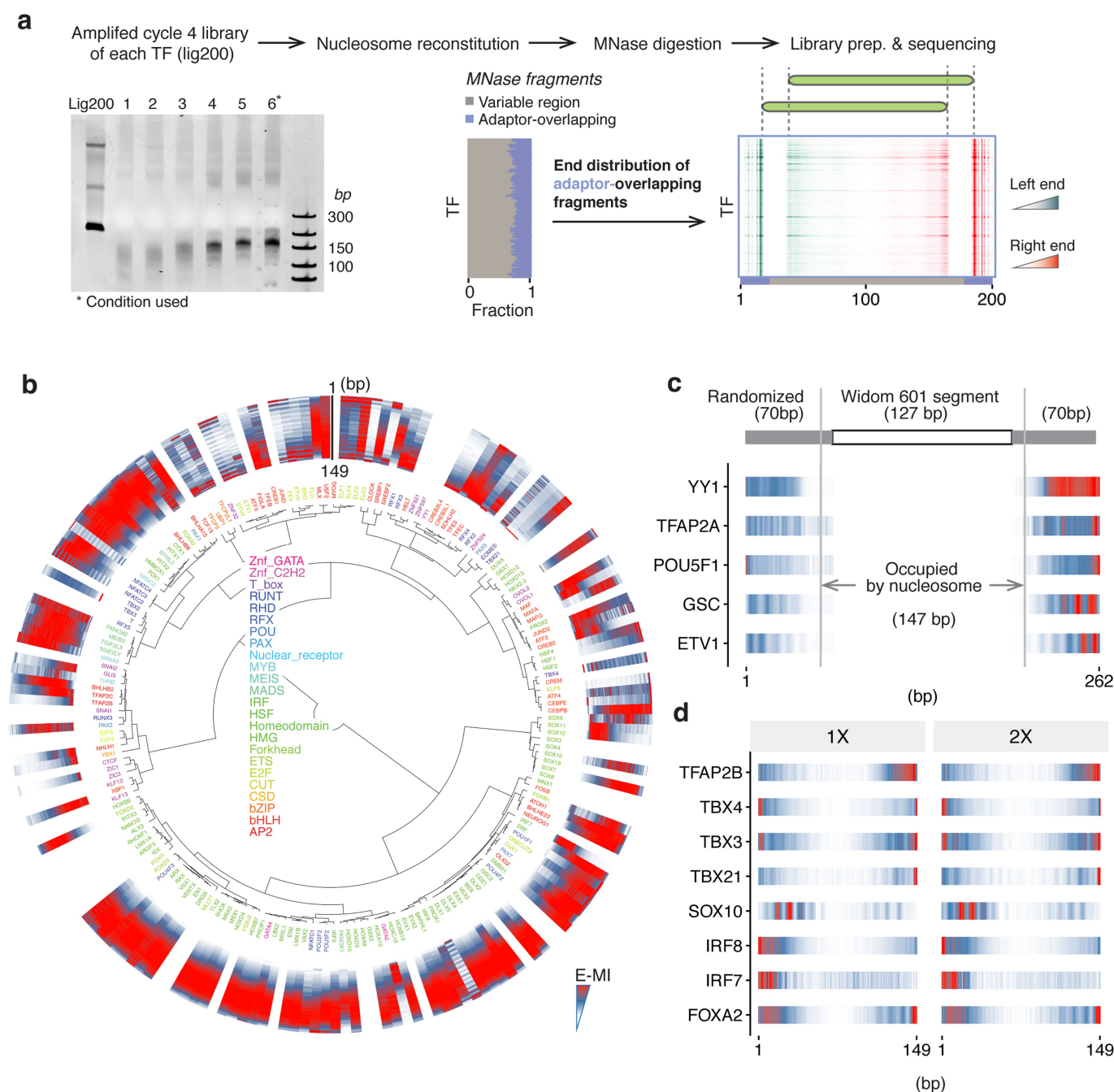
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | NCAP-SELEX with lig200. **a**, Hierarchical clustering of the E-MI diagonals for NCAP-SELEX with the 200-bp ligand (lig200). The E-MI diagonal for each TF is oriented radially. The randomized region is 154 bp and contains 149 windows for MI calculation between neighbouring 3-mers. The names of the TFs are coloured by family with the colouring scheme indicated on the centre. TFs from the same family tend to be clustered together (for example, SOX, indicated). Because of the gradient of nucleosome occupancy, the penetration of the E-MI signal into the centre of the E-MI diagonals (E-MI penetration; see Supplementary Methods for details) reflects the ability of each TF to bind to nucleosomal DNA. Note that almost all TFs have lower E-MI towards the centre of lig200, indicating their lower affinity to nucleosomal DNA than to free DNA. The decrease of E-MI towards the centre is rarely observed in the absence of the nucleosome. Note that the binding inhibition of TF to nucleosomal DNA occurs in the absence of higher-order effects, such as chromatin compaction, remodelling or histone modification. This result directly verifies the mutually antagonistic role of TFs and the nucleosome^{13,43,44}, which has been biochemically validated in only a few cases^{45,46}. The E-MI diagonals shown are scaled for each TF (see Supplementary Methods). Owing to the fixed adaptor sequences, TFs may prefer one end of the lig200 over the other end. **b**, E-MI penetration of individual TFs on lig200. TFs are ordered according to their E-MI penetration depth towards the centre of the ligand. This order reflects the ability of TFs to bind nucleosome-occupied DNA. Note that the penetration of E-MI into the ligand centre (E-MI penetration; see Supplementary Methods for details) varies strongly between the TFs. TFs representing either of the two ends are coloured red and exemplified in **c**. **c**, The diagonal of E-MI for TFs with high (above dotted line) and low (below dotted line) E-MI penetrations. Because HT (blue) and NCAP-SELEX (black) may differ in stringency, each E-MI diagonal is normalized by dividing its maximum value. On lig200 the central 94 bp (shaded grey) is always occupied by a nucleosome. **d**, Correlation between E-MI penetration and the capability of TFs to bind nucleosomal DNA in vivo. Per base-pair coverage of MNase fragments (>140 bp) at ChIP-seq peaks of the TFs (*x* axis) is plotted against their E-MI penetration (*y* axis) in NCAP-SELEX. The calculation of Pearson's *r* and the correlation test is performed for *n* = 20 TFs. The observed correlation suggests that the ability of TFs to bind nucleosomal DNA in

NCAP-SELEX (E-MI penetration) partially explains the nucleosome occupancy at the sites of TFs in K562 cells. Thus the biochemical ability of TFs to bind to nucleosomal DNA also affects their binding in vivo. **e**, Left, E-MI heat map of T (brachyury) in HT-SELEX using lig200. Pairwise E-MI for all 3-mer pairs is presented as a heat map. The signal is only visible near the diagonal, no E-MI signal is detected across approximately 80 bp. Right, the gyre-spanning mode (arrow) of T (brachyury) on lig147. The corresponding motif is derived with the indicated seed for a specific position (number in the parentheses) in the high E-MI region (arrow). Position weight matrix generation follows our previous method⁴⁷ using multinomial 1. **f**, Type 2 binding of Brachyury (T) stabilizes nucleosome from dissociation. log₂ ratio of E-MI between the bound and unbound libraries (cycle 5) is calculated for both the type 2 binding and for the background E-MI level (see Supplementary Methods for details) of Brachyury (T). Compared to the unbound, the bound library has stronger type 2 binding but a similar background. As a control, for 20 random TFs (Rnd), the log₂ ratio of E-MI between the bound and unbound libraries is also calculated for both the type 2 binding (hypothetic) and for the background E-MI level. For these TFs the bound libraries have similar E-MI strength as the unbound in the region corresponding to the type 2 binding of Brachyury (T). Data are mean ± s.d.; two-sided *t*-test was used, 95% confidence intervals, 0.097 – 0.202 (T) and –0.008 – 0.004 (random TFs). The sample sizes are *n* = 20 libraries for random TFs and *n* = 4 independent SELEX replicates for Brachyury (T). The raw data for the random control TFs are listed in Supplementary Data 3. **g**, E-MI heat map of TBX2, ETV4 and ETV1 in NCAP-SELEX using lig147. The E-MI signals across approximately 80 (type 2) or 40 bp (type 1) are indicated with arrows. The corresponding motif of each binding type is derived with the indicated seed for a specific position (number in the parentheses) in the high E-MI regions (arrows). Note that the E-MI signals across approximately 40 bp are position-specific, with one binding event being observed near the dyad, and the other(s) on the opposite side of the nucleosome, with the two contacts separated by approximately 180°. This binding mode can be achieved by TF dimers that contact nucleosomal DNA in a pincer-like manner. However, as the individual TFs are located far from each other in this binding mode, it probably suggests that the nucleosome may have two allosteric states, or may form a higher-order complex with these TFs.

43. Ramachandran, S. & Henikoff, S. Transcriptional regulators compete with nucleosomes post-replication. *Cell* **165**, 580–592 (2016).
44. Li, M. et al. Dynamic regulation of transcription factors by nucleosome remodeling. *eLife* **4**, e06249 (2015).
45. Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. & Zaret, K. S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev.* **23**, 804–809 (2009).

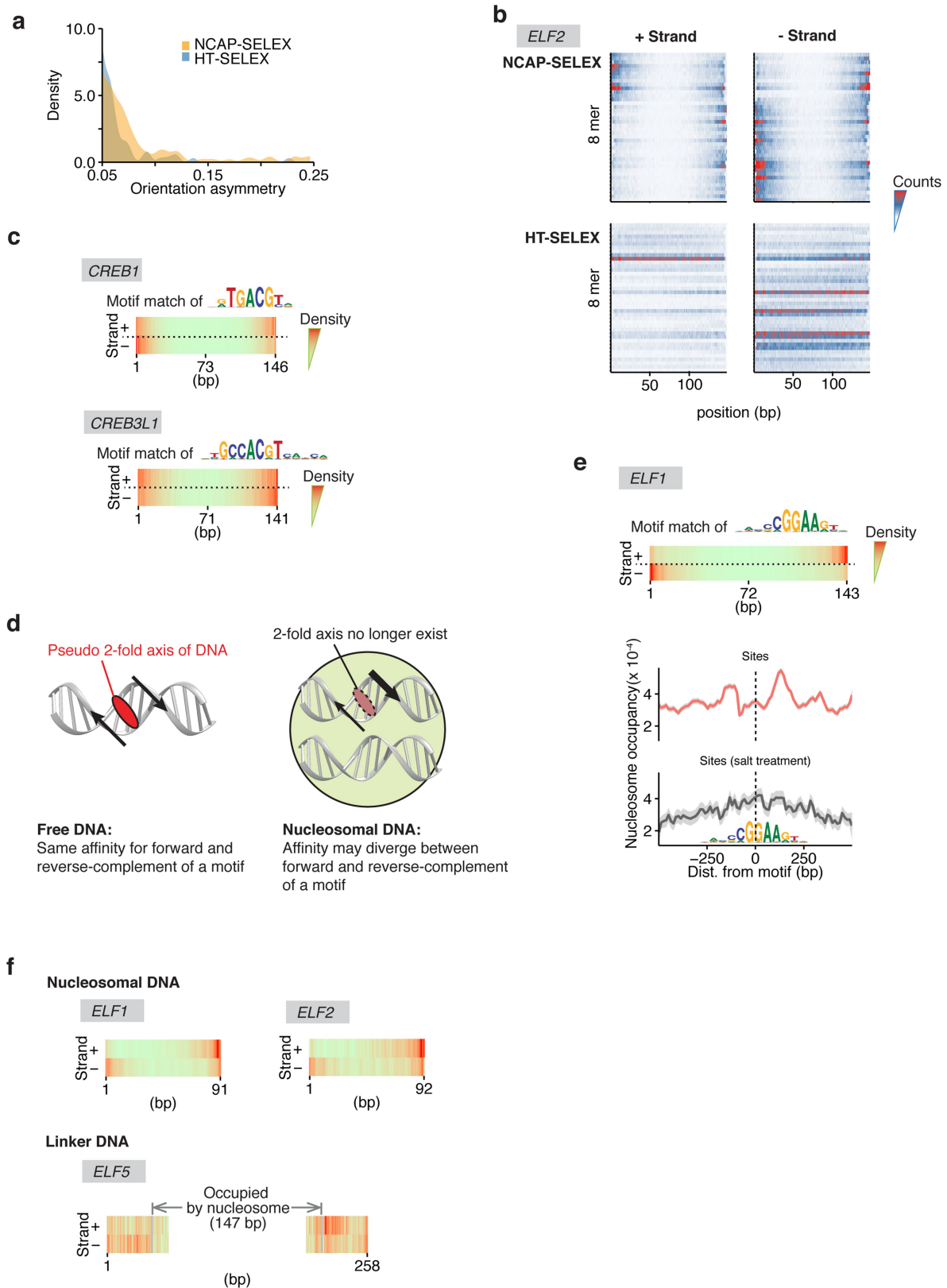
46. Hayes, J. J. & Wolffe, A. P. Histones H2A/H2B inhibit the interaction of transcription factor IIIA with the *Xenopus borealis* somatic 5S RNA gene in a nucleosome. *Proc. Natl Acad. Sci. USA* **89**, 1229–1233 (1992).
47. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).



Extended Data Fig. 3 | Control experiments with lig200.

a, Determination of nucleosome positions for NCAP-SELEX libraries (lig200, all TFs). To examine if nucleosome has preferred positioning on lig200, nucleosomes were loaded onto the amplified cycle 4 NCAP-SELEX library of each TF. After digestion with MNase, the remaining DNA fragments were collected and sequenced. A titration was first carried out to find the appropriate concentration of MNase. As shown in the gel image (left, see Supplementary Fig. 1 for gel source image), 4.8, 2.4, 1.2, 0.6, 0.3, 0.15 U of MNase (lane 1–6) were added into each 25- μ l reaction containing the purified nucleosome. According to the results, the condition marked by an asterisk was chosen for the reactions to determine nucleosome position. After sequencing, the fractions of MNase fragments that mapped to the variable region (grey) and to the adaptor-overlapping region (blue) of lig200 are visualized (middle, each row corresponds to a TF). To identify potential positional preference of nucleosome on lig200, the adaptor-overlapping fragments are analysed for their end distributions. Distributions of both the left end (cyan) and the right end (red) of the MNase-digested fragments on lig200 are shown (right, each row corresponds to a TF). Such distributions likely indicate that nucleosomes have two relatively preferred positions on lig200 (illustrated by cartoon in green). Note that most nucleosomes are not positioned by

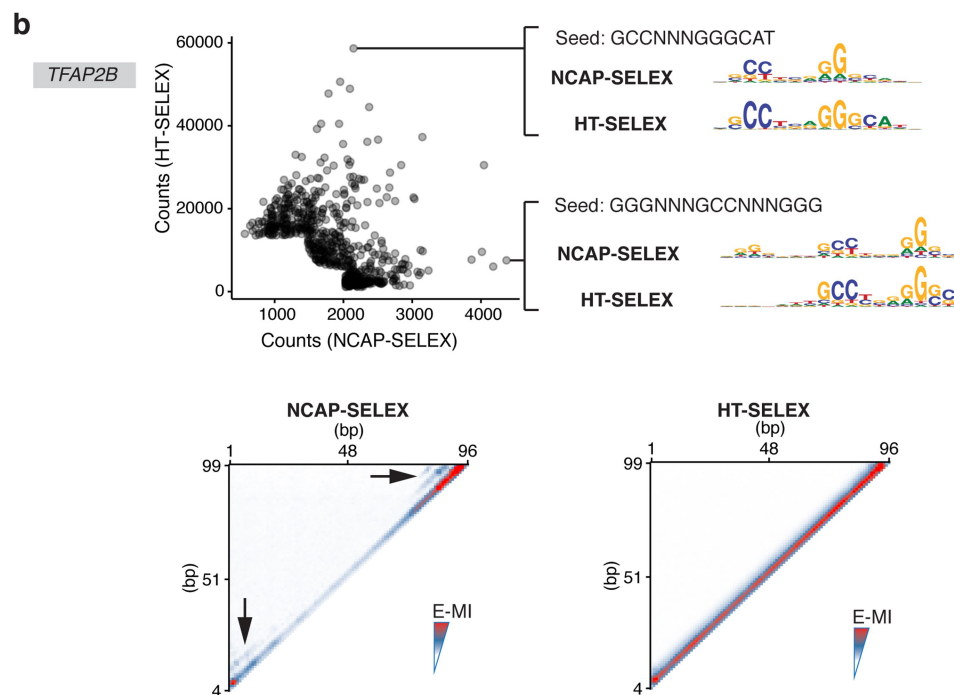
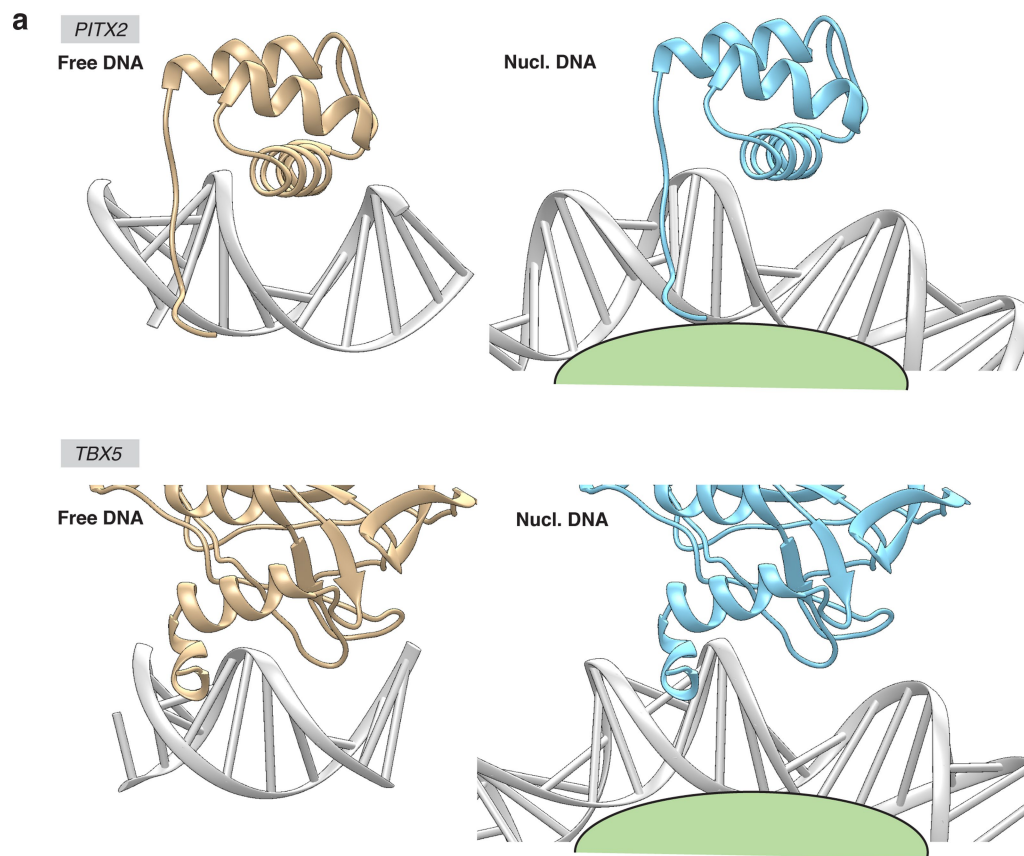
the adaptor (middle) thus are randomly distributed. **b**, E-MI diagonals for HT-SELEX with the 200-bp ligand (lig200). TFs are arranged according to the clustering for NCAP-SELEX libraries (Extended Data Fig. 2a) to facilitate comparison. TFs without a lig200 HT-SELEX control are left as blank. The E-MI diagonal for each TF is oriented radially and the names of the TFs are coloured by family as indicated. The E-MI diagonals are scaled for each TF. Some TFs show preferred positions on lig200, probably due to the fixed adaptors. **c**, TFs prefer free DNA to the edge of a nucleosome. For a few randomly chosen TFs, NCAP-SELEX was run using a ligand (Lig70Nlinker, sequence in Supplementary Table 2) that positions nucleosome at its centre by embedding a segment of Widom 601 sequence, and with randomized flankings. At a low resolution, the E-MI signal of TFs decreases monotonically towards the nucleosome-occupied region. Thus the higher E-MI at the flankings of lig200 (Extended Data Fig. 2a) suggests the preference of TFs for free DNA, rather than for the edge of a nucleosome. E-MI diagonals are scaled for each TF. **d**, E-MI diagonals for TFs at doubled concentrations. The concentration effect on the E-MI diagonal of TFs is explored by running NCAP-SELEX at doubled ($2\times$) concentrations for a few randomly chosen TFs. Compared to the E-MI diagonal with the original TF concentrations ($1\times$), the change in the E-MI pattern is minor.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Nucleosome breaks the rotational symmetry of DNA. **a**, Density plot representing the orientational asymmetry of all TFs in NCAP-SELEX and in HT-SELEX. In NCAP-SELEX, more TFs bind with high orientational asymmetry than in HT-SELEX. A few TFs can also prefer different ends of the ligand for the two binding directions in HT-SELEX; this is likely induced by the adaptor sequences. However, there are more TFs with higher orientational asymmetry in NCAP-SELEX libraries, despite the fact that for most TFs their signals are stronger in HT-SELEX libraries. **b**, Orientation asymmetry of ELF2 revealed by using top 8-mers. Each row of the heat map corresponds to the counts distribution of a top 8-mer (non-palindromic) across the positions of the SELEX ligand. Hits of the top 8-mers occur at different ends for different strands of nucleosomal DNA (that is, an 8-mer and its reverse-complement prefer different ends), whereas their distribution is relatively homogeneous for free DNA. **c**, Orientation asymmetry of CREB TFs. CREB TFs have different motif density distributions for the two strands of nucleosomal DNA. The motif used for matching is indicated above. The minus strand profile is from the density of the reverse-complement motif. **d**, Break of the two-fold rotational symmetry of DNA induces preferred orientation of TFs. Left, free DNA has a pseudo-two-fold axis (red ellipse) perpendicular to the helix axis. Motifs in two orientations are symmetric with each other

with respect to a 180° rotation centred on the axis. Right, for motifs on nucleosomal DNA, if the other strand of DNA or the histone proteins (green) affect binding, the two-fold axis of DNA no longer exists, as a 180° rotation centred on the axis no longer generates an identical conformation (the rotated image not superimposable with the original one). The break of rotational symmetry occurs also on the linker DNA that immediately flanks the nucleosome (**f**). **e**, Top, the orientational asymmetry of ELF1 in NCAP-SELEX of lig200. Bottom, the asymmetric nucleosome distribution around genomic ELF1 sites (top). Such asymmetry is not observed for the same ELF1 sites after a 30 min 500 mM KCl treatment to mobilize the nucleosome (bottom). ELF1 motif matches are positioned at the centre. Frequency of the centre of MNase-fragments (140–170 bp) is visualized for nearby regions to represent the nucleosome occupancy. Each profile ($n = 999$ data points) is LOESS smoothed with a span of 0.05 and the shaded band indicates the s.e.m. **f**, The orientational binding of ELF occurs on both the nucleosomal DNA and the nearby linker region. The motif matches of ELF on lig147 (top) suggest that the orientational binding occurs on nucleosomal DNA. In addition, the motif matches of ELF on the 293-bp ligand (bottom; nucleosome positioned at the centre, ligand schematic in Extended Data Fig. 3c) indicates that the orientational binding also occurs on nearby linker DNA regions.



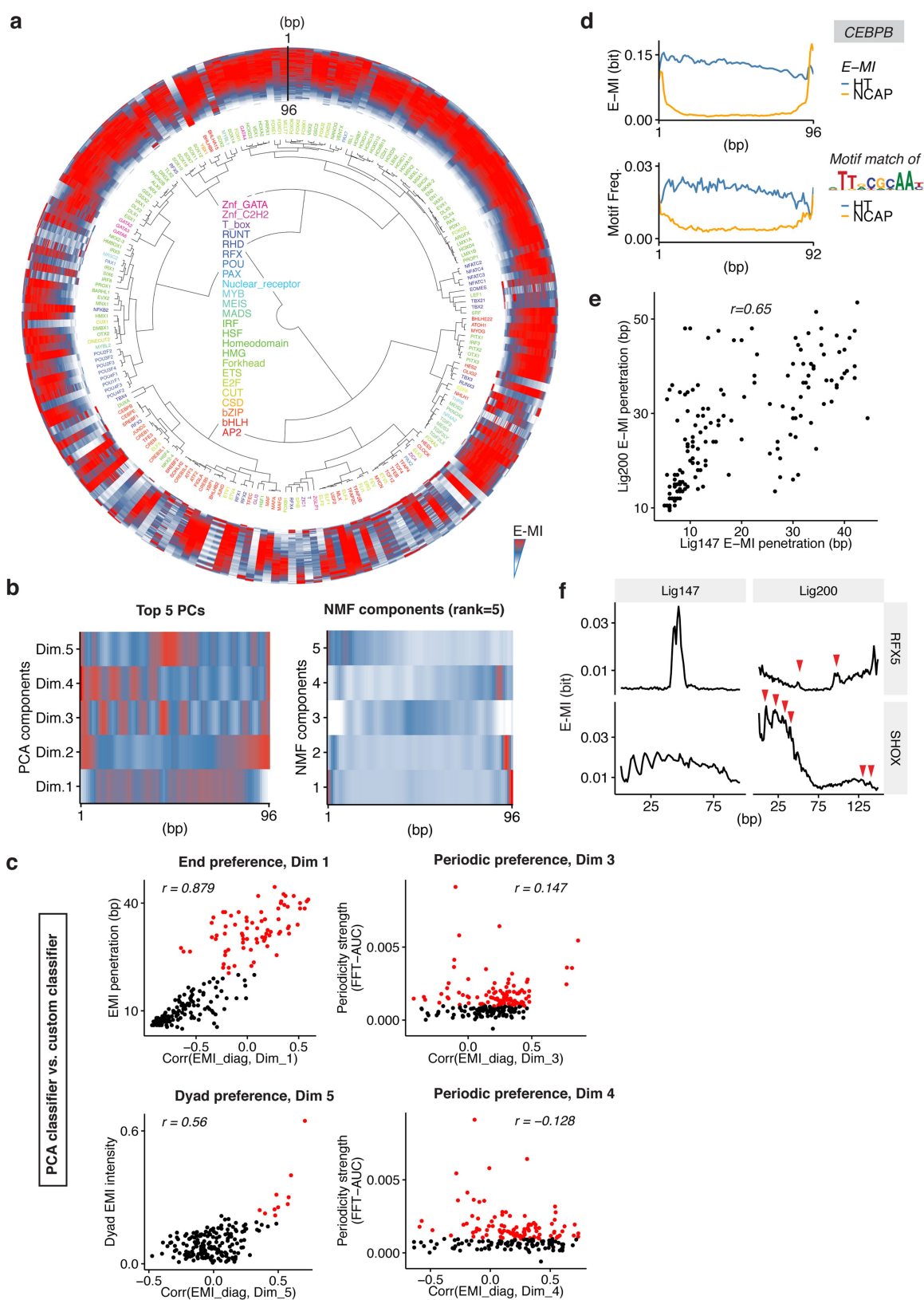
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | TFs can bind nucleosomal DNA without substantial motif change. **a**, Cartoons showing that TFs are theoretically able to contact grooves of the bent nucleosomal DNA from the solvent-exposed side. The left panel for each TF shows the structures (Protein Data Bank (PDB); PITX2: 2LKX, TBX5: 2X6V). For the right panels of each TF, the PDB structure of the TF is aligned to the nucleosome structure (3UT9) as described in the Supplementary Methods (section 'FFT analysis and structure alignment'). The corresponding base pairs of the nucleosomal DNA were replaced with Coot⁴⁸ according to the DNA sequence in the PDB structure of each TF. The models are visualized with UCSF Chimera⁴⁹. **b**, TFAP binds nucleosomal DNA with slightly

different specificity than free DNA. The scatter plot (top) shows the counts of gapped 9-mers from SELEX libraries of TFAP2B, enriched with NCAP-SELEX (x axis) and HT-SELEX (y axis). The examined 9-mers consist of three segments of trimers interspaced with two gaps (0–5 bp). Only the most enriched 9-mers (top 300 in each library and in the combined library) are shown for clarity. For comparison, the most differentially enriched gapped 9-mers were also used as seeds to derive the corresponding motifs from both libraries (right). The heat map (bottom) shows the pairwise E-MI for all combinations of positions on lig147, in the presence (left) and absence (right) of nucleosome. The arrowheads indicate the additional signals developed in the presence of nucleosome.

48. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).

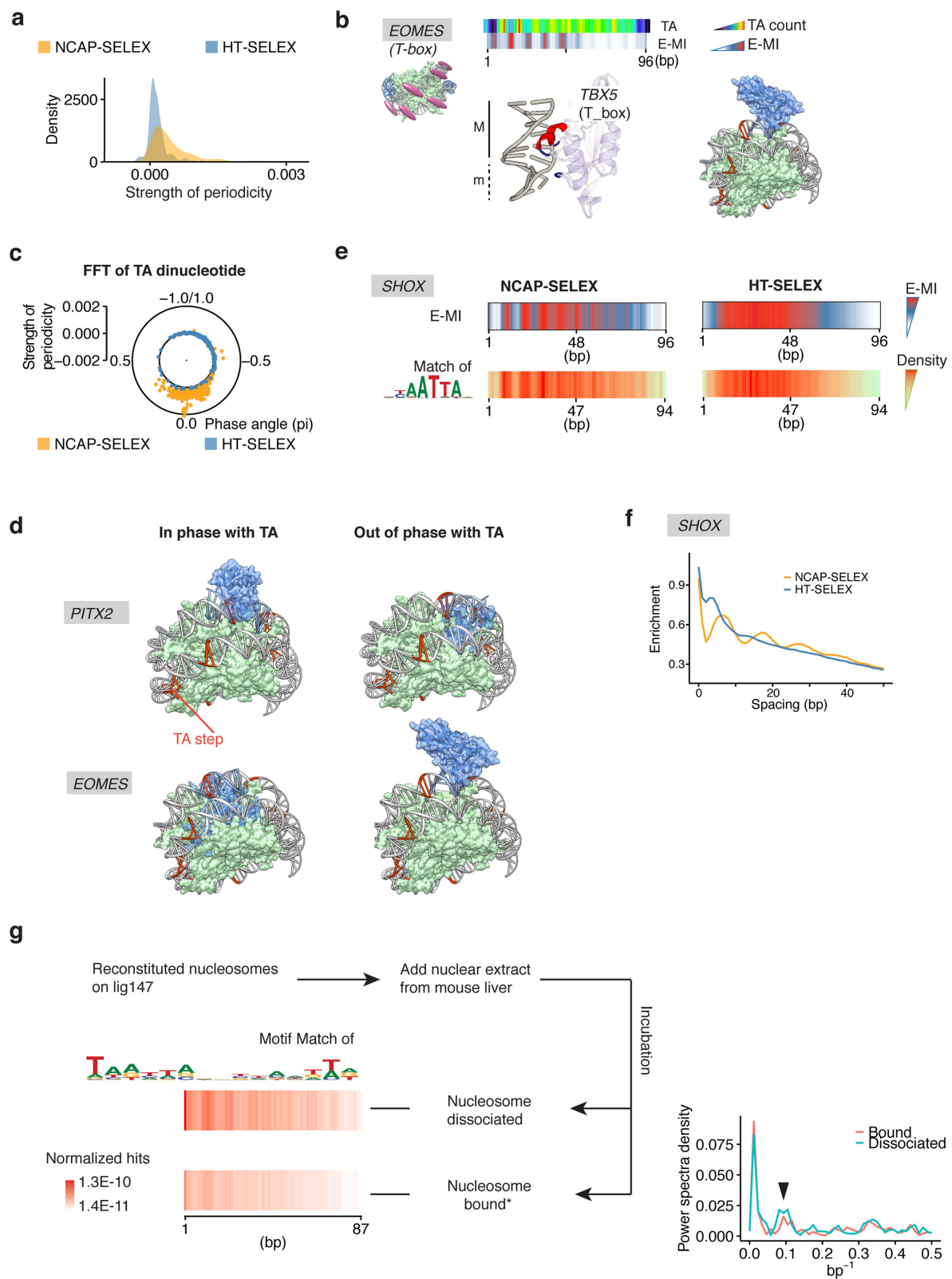
49. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | NCAP-SELEX with lig147. **a**, E-MI diagonals for HT-SELEX with the 147-bp ligand (lig147). TFs are arranged according to the clustering for NCAP-SELEX libraries (Fig. 3a) to facilitate comparison. The E-MI diagonal for each TF is oriented radially and scaled. The names of the TFs are coloured by family as indicated. **b**, The top five principal components (PCs) and the components from non-negative matrix factorization (NMF) with rank equal to five. The E-MI diagonals of lig147 ($n = 195$ TFs) were used in the dimension reduction. For visualization purposes, each component is centred and scaled. Note that the five principal components (left) correspond well to the three identified positional preferences of TFs on nucleosomal DNA (end: dim 1, 2; periodic: dim 3, 4; dyad: dim 5). **c**, Comparison between the scores from principal-component classifiers and custom classifiers. Red points indicate the TFs defined as displaying respective preferences according to custom classifiers. The PC classifiers are well in accordance with custom classifiers for the end and the dyad preferences (left), but not for the periodic preference (right). Because the phase of periodic preference can vary continuously whereas principal components can only capture discrete values, the custom FFT-based classifier is more natural for such purposes. The libraries of $n = 195$ TFs were used in the analyses. The correlation

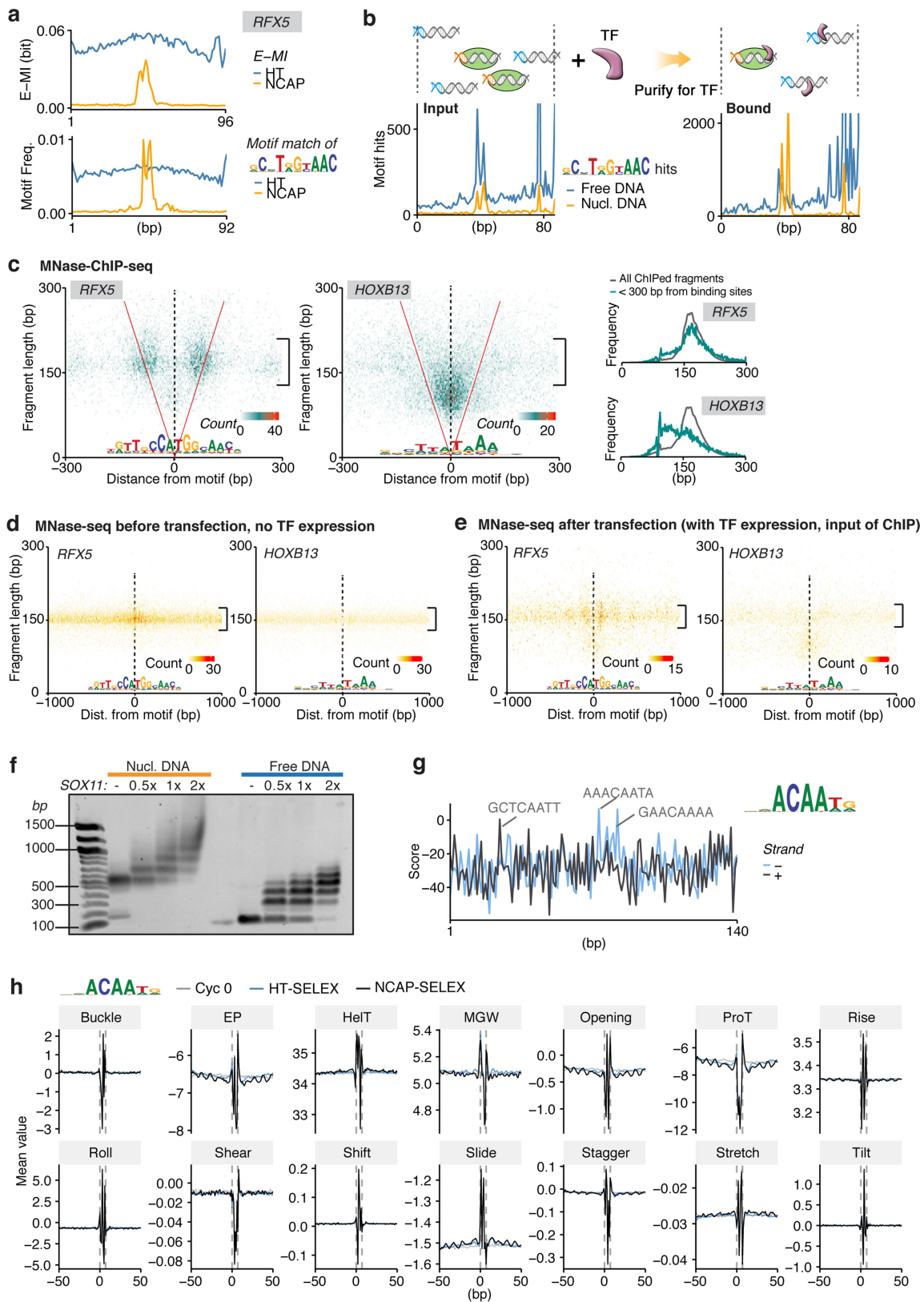
coefficients (Pearson's r) are also indicated. **d**, E-MI diagonal and motif-matching results for the bZIP factor CEBPB. In HT-SELEX (without nucleosome), the binding signal is more distributed across the ligand. **e**, Pearson's correlation between the E-MI penetrations of TFs on lig200 and on lig147. The libraries of $n = 155$ TFs, which are successful with both lig200 and lig147, were used in this analysis. The end preference of TFs on lig200 reveals that they prefer free DNA to nucleosomal DNA. The free-DNA preference also probably explains the end preference of TFs on lig147 owing to the observed correlation of E-MI penetrations. For each TF, the E-MI penetration values differ between lig147 and lig200 because free-DNA regions are expected near the ends of lig200, but not present on lig147. **f**, Correspondence between the E-MI patterns of TFs on lig147 and on lig200. The E-MI diagonals of RFX5 and SHOX on lig200 and those on lig147 are plotted together for comparison. The peaks on lig200 that illustrate the central preference of RFX5 and periodic preference of SHOX are indicated with red arrowheads. The weaker preference patterns on lig200 are due to the delocalization of the nucleosome on lig200, however they are still visible because the two fixed adaptors dictate two weakly preferred nucleosome positions.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | TFs with periodic preferences. **a**, Density plot showing the periodicity strength of all TFs in NCAP-SELEX (orange) and HT-SELEX (blue). Note that the overall periodicity of E-MI is stronger for the NCAP-SELEX library compared to the free-DNA HT-SELEX library. **b**, A minor-groove binder prefers exposed minor grooves (m) on nucleosomal DNA. The E-MI diagonal of EOMES (T-box) is out of phase with the TA dinucleotide peaks, suggesting that it binds positions where the minor groove of nucleosomal DNA is facing outside (TA peaks indicate nucleosome–DNA contacts, whereas E-MI visualizes TF–DNA contacts, see Supplementary Methods for details). Accordingly, the TBX5 (T-box) structure (PDB entry 2X6V) shows contacts with DNA principally in the minor groove. Cartoon representation to the right shows that the steric hindrance is minimal when TBX5 (blue) binds out of phase with TA (orange) on the nucleosome structure (PDB entry 3UT9). **c**, Strength and phase of the approximately 10-bp periodicity of the TA dinucleotide in NCAP-SELEX and HT-SELEX libraries. For the library (lig147) enriched by a specific TF, the strength and phase information is derived from FFT of the TA counts at each position of the library. In the polar plot, each dot represents the library of one TF. The overall periodicity is stronger in the NCAP-SELEX libraries (yellow) than in the HT-SELEX libraries (blue), suggesting an enrichment of nucleosome signal. The TA phases in the NCAP-SELEX libraries of all TFs are similar, thus the rotational positioning of nucleosome on the SELEX ligand is similar for the libraries of all TFs. By contrast, the phase of the E-MI periodicity is much more dispersed (Fig. 4a), suggesting the preference of TFs towards different grooves of DNA. **d**, Cartoon representations of the 3D structures of PITX2 (PDB entry 2LKX) and TBX5 (T-box, PDB entry 2X6V) in complex with nucleosomal DNA. TBX5 structures were shown to illustrate the groove preferences of EOMES (T-box). The DNA ligand

in the nucleosome structure (PDB entry 3UT9) contains phased TA steps (orange). Consistent with the SELEX result, PITX is more compatible with nucleosomal DNA when it binds in phase with TA, whereas T-box is more compatible when it binds out of phase with TA. Therefore, when a TF binds nucleosomal DNA according to the identified patterns, the steric conflict between TF and the histones is minimized. **e**, E-MI diagonal and motif-matching results for SHOX in NCAP-SELEX and HT-SELEX. The E-MI diagonal agrees with the motif-matching result. **f**, The approximately 10-bp periodicity for the preferred spacing of SHOX dimers on nucleosomal DNA. In NCAP-SELEX libraries of many periodic binders (SHOX as an example), enrichment of the most abundant 3-mer tandem repeats oscillates as a function of the spacing between the repeats. The enrichment is evaluated by the log2 ratio between the observed and expected occurrences. The observed approximately 10-bp periodicity with dimer spacing originates from the periodic availability of nucleosomal DNA. However, in most cases binding appears not to be cooperative, on the basis of the fact that the observed frequency of ligands with two motifs can be well estimated by the frequency of ligands that contain only one motif (data not shown). **g**, Homeodomain TFs from mouse liver prefer periodic positions on nucleosomal DNA. Motif hits of homeodomain TFs show a periodic pattern for both the nucleosome-bound and nucleosome-dissociated (unbound) libraries after incubation with mouse liver nuclear extract; however, the unbound library has more motif hits, indicating that binding events to the presented motif facilitate the dissociation of nucleosome. To more clearly visualize the approximately 10-bp periodicity, the Fourier-transformed spectra for both libraries are also shown to the right. The arrowhead indicates the peaks for the approximately 10-bp periodicity.



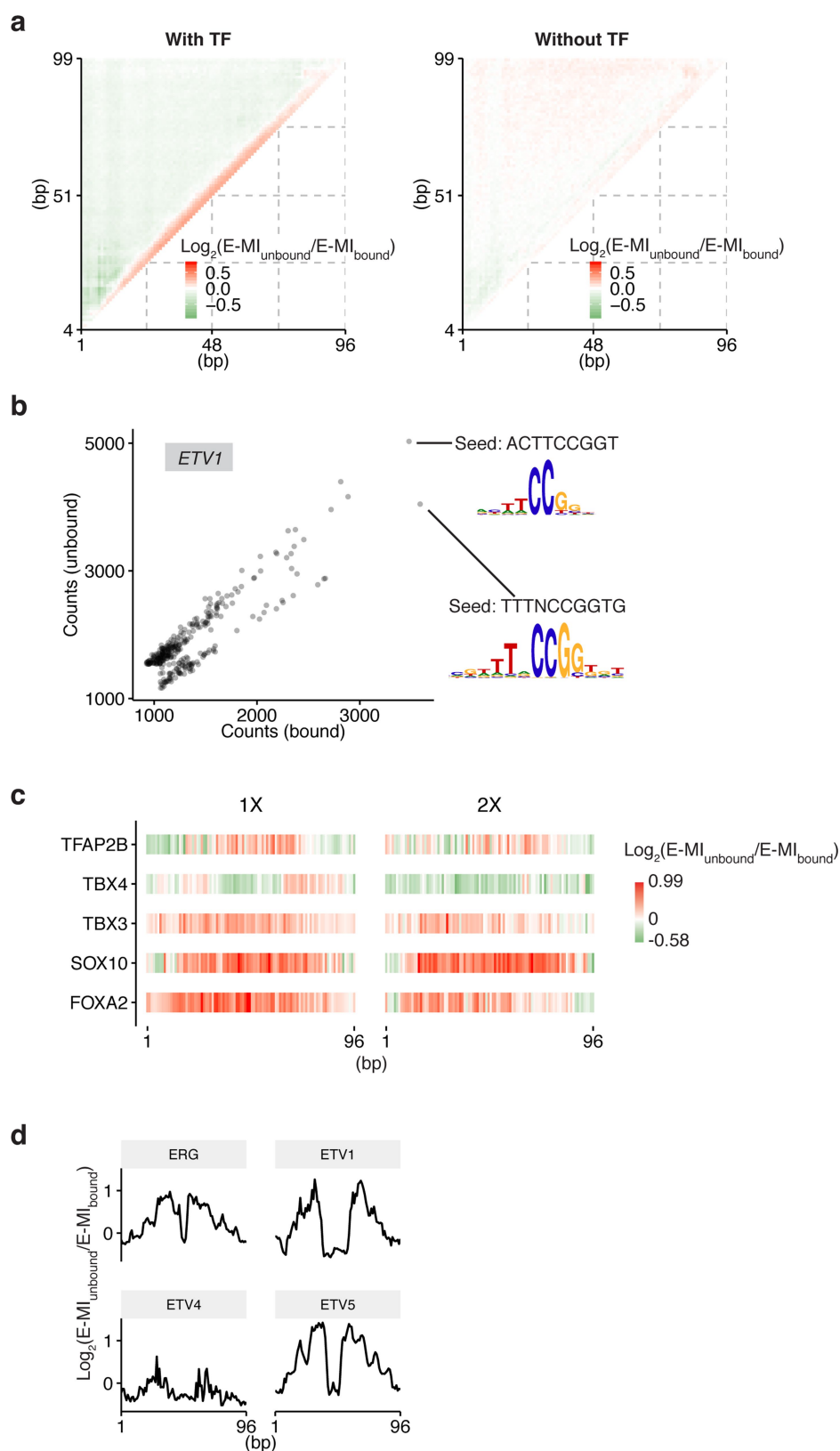
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | TFs with the dyad preference. **a**, E-MI diagonal and motif-matching results for RFX5. The distribution of binding events is more spread in the absence of nucleosome (HT-SELEX). **b**, The design of the competition assay and the raw counts of RFX5 motif matches. Differently barcoded nucleosomal DNA (orange) and free DNA (blue) were mixed as input, and incubated with the TF protein. Purification for the TF-bound species was then performed. Matches of the indicated RFX5 motif was counted for both the nucleosomal DNA (orange) and the free DNA (blue), and for both of the input and the bound libraries. On nucleosomal DNA, more motif hits near the centre of the ligand are observed after purification. **c**, MNase–ChIP fragments near the binding sites of RFX5 and HOXB13. Motif matches within MNase–ChIP peaks of each TF are positioned at the centre. Counts of MNase–ChIP fragments are binned to 3 bp by 3 bp bins according to their lengths and centre positions. Nucleosome distribution is reflected by the signal intensity of the approximately 150-bp fragments (bracket). This visualization resembles the reported ‘V-plot’⁵⁰. Length distribution of all ChIP fragments and that of fragments <300 bp from the TF sites are shown on the right. Note that HOXB13 enriches ChIP fragments of approximately 120 bp at its sites (middle), suggesting that, similarly to most TFs^{50,51}, its binding sites in the genome are depleted of nucleosomes. By contrast, RFX5 enriches nucleosome-sized fragments (left). Most of the enriched fragments also have their centre positioned between the red ‘V’ lines, and thus overlap with the TF motifs. **d**, Nucleosome distribution

near the binding sites of RFX5 and HOXB13 before transfection (no TF expression). MNase–seq fragments around the identified TF sites are visualized as in **c**. The sites later bound by exogenous RFX5 are located at the maximum of nucleosome occupancy (left). **e**, Nucleosome distribution near the binding sites of RFX5 and HOXB13 after transfection (with TF expression). The nucleosomes are now positioned beside the exogenous RFX5 sites (left). **f**, EMSA of SOX11 complexes with nucleosome and with free DNA. Nucleosome is reconstituted and purified using a modified Widom 601 sequence, which contains a SOX11 binding sequence (extracted from cycle 4 SELEX library) embedded close to the dyad. Each 40 µl reaction contains 1 µg DNA, together with SOX11 protein at a molar ratio of 0, 0.5, 1, 2 (indicated at the top of each lane) to DNA. Here the observed multiple shifts probably reveal the binding of SOX11 to additional weaker sites on the ligand (shown in **g**). For gel source data, see Supplementary Fig. 1. **g**, The score of SOX11 motif across the EMSA ligand (see Supplementary Methods for ligand sequence). The top three binding sites are indicated. **h**, DNA shape features around SOX11 motifs. DNA shape features were calculated using DNASHapeR^{52,53}, for NCAP–SELEX (black), HT–SELEX (blue), and cycle 0 (input, grey) libraries. The black line is plotted last thus may hide other lines when all values are similar. The boundary of each motif is indicated with dashed vertical lines. Only the ligands with motifs around the centre (position range: 36–58) are included in the analysis.

50. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA* **108**, 18318–18323 (2011).
51. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203–209 (2014).

52. Chiu, T. P. et al. DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211–1213 (2016).
53. Chiu, T. P., Rao, S., Mann, R. S., Honig, B. & Rohs, R. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res.* **45**, 12565–12576 (2017).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | TF binding affects the stability of nucleosome.

a, E-MI difference between the bound and the unbound cycle 5 libraries. The bound and the unbound libraries were collected either in the presence (left) or in the absence (right) of TFs. The heat maps visualize E-MI differences between the bound and unbound libraries for all position combinations of 3-mer pairs, and each pixel on the heat map is a mean of the E-MI difference of all the examined TFs at this pixel. For individual TFs, the value at each pixel is calculated as $\log_2(E-MI_{\text{unbound}}/E-MI_{\text{bound}})$. Testing nucleosome dissociation in the absence of the TF aimed to verify whether the TF motifs on lig147 by themselves can affect the stability of the nucleosome. Note that in general, binding events close to the centre of nucleosomal DNA more efficiently dissociated the nucleosome (left). This observation is in accordance with the mutually exclusive nature between TFs and the nucleosome. Although TFs generally have lower affinity to the centre of the lig147, it is also conceivable that TF binding close to the centre will more efficiently undermine the DNA–histone interactions, and in turn lead to a higher rate of nucleosome dissociation. TFs bound

close to the ends could have decreased the flexibility of the DNA there and subsequently disfavour the dissociation of DNA ends from the histones, which in turn contributes to nucleosome stability. **b**, The efficiency of nucleosome dissociation induced by ETV1 is dependent on its binding specificity. To displace nucleosome, binding with the shorter motif is more efficient than binding with the longer motif, because the shorter motif is more enriched in the dissociated library (unbound). **c**, Differential E-MI diagonals for TFs at doubled concentrations. The ability of each TF to dissociate or stabilize nucleosome is revealed by the log ratio of E-MI between the unbound and the bound cycle 5 libraries (differential E-MI). The concentration effect on the differential E-MI diagonal of TFs is explored by running NCAP–SELEX followed by the dissociation assay at doubled ($2\times$) concentrations of the TFs. The differential E-MI diagonals at $2\times$ TF concentrations resemble those at the original ($1\times$) TF concentrations. **d**, Differential E-MI diagonals for the four ETS family TFs indicated by asterisks in Fig. 5a.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Modes of TF–nucleosome interaction.

a, For each TF, the strengths of all identified TF–nucleosome interaction modes, together with its ability to dissociate nucleosome, are shown in the heat map. The displayed features include the positional preference of each TF (E, end; P, periodic; D, dyad) on nucleosomal DNA, gyre-spanning binding mode (Gs), orientational asymmetry (Asym), and the ability of each TF to dissociate nucleosome (Ds). TFs succeeding only in NCAP–SELEX with lig200 are presented to the right for their orientational asymmetry. In the heat map values are scaled into 0 to 1 for each mode, except for the dissociation, in which TFs that stabilize nucleosome are given negative values (green). The raw data are provided in Supplementary Table 5. **b**, All the identified modes can be explained by the structural features of nucleosome. TFs with the end preference (E) bind nucleosomal DNA close to the entry and exit positions. This preference is in line with the probability of spontaneous dissociation (breathing) of nucleosomal DNA, which decreases from the end to the centre^{54–56}. TFs with a strong end preference are likely less compatible with nucleosomal DNA thus only bind to the dissociated regions. These TFs could be structurally hindered by nucleosome, because one side of the nucleosomal DNA is masked by the histones. Moreover, nucleosomal DNA is bent sharply, which could impair TF–DNA contacts if TFs have evolved to specifically bind to free DNA. TFs with the periodic preference (P) binds approximately every 10.2 bp positions on nucleosomal DNA. This preference arises also because nucleosomal DNA is accessible only from one side, which leads to significant accessibility change along each pitch (approximately 10.2 bp) of the DNA helix. TFs that bind to short motifs, or to discontinuous motifs, are still able to occupy the available periodic positions on nucleosomal DNA. TFs with the dyad preference (D) tend to bind close to the nucleosomal dyad. Structurally, the dyad is distinct from other regions of the nucleosomal DNA. The dyad contains only a single DNA gyre, and features the thinnest histone disk^{29,37}. These characteristics of the dyad DNA reduce the steric barrier for TF binding. The relatively weak DNA–histone interaction around the dyad could allow TFs that bend DNA upon binding (for example, SOXs⁵⁷) to deform DNA more easily at the dyad compared to other positions. In addition,

the entry and exit of nucleosomal DNA are also close to the dyad; together with the dyad DNA, they provide a scaffold for specific configurations of TFs. FOXA has been suggested to make use of this scaffold to achieve highly specific positioning close to the dyad^{39,58}. However, the dyad positioning of FOXA is not observed in this study using eDBD, potentially because the full length of FOXA is required for its interaction with the nucleosome⁵⁹. A few T-box TFs were found to bind nucleosomal DNA with the gyre-spanning binding mode (Gs). This mode is observed because DNA grooves align across the two nucleosomal DNA gyres²⁹. The parallel gyres could specifically associate with TF dimers, or TFs with long recognition helices or multiple DNA-binding domains. The dual-gyre binding is possible only on nucleosomal DNA, and it thus stabilizes the nucleosome from dissociation, and may therefore function to lock a nucleosome in place at a specific position. Many TFs such as ETS and CREB show an orientational asymmetry (Asym) upon binding to the nucleosomal DNA. The nucleosomal environment has induced such preference by breaking the local rotational symmetry of DNA. In accordance with the mutually exclusive nature of TF and nucleosome binding, most TFs were found to dissociate nucleosomes (Ds). While nucleosome weakens the affinity of incompatible TFs, binding of such TFs are expected to weaken the nucleosome–DNA contacts as well. The ability of TFs to dissociate nucleosome is required for them to open chromatin and to activate transcription. Moreover, we also observed TFs that both stabilize and destabilize nucleosomal DNA, depending on their relative position of binding. Such ability could be used to more precisely position local nucleosomes. All the identified TF–nucleosome interactions suggest that the TF–nucleosome interaction could be more complicated than the previously suggested pioneer/non-pioneer classification of TFs¹¹. We observed that for eDBD of almost all TFs, including known pioneer factors such as FOX and SOX, free DNA was nonetheless preferred over nucleosomal DNA. However, some pioneer factors can bind relatively better to the interior of the nucleosome (for example, FOX and SOX). In addition, some other TFs prefer nucleosomal DNA at restricted positions, or with one of their multiple binding motifs. These strategies are likely related to the access of pioneer factors to nucleosomal DNA.

54. Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
55. Anderson, J. D. & Widom, J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J. Mol. Biol.* **296**, 979–987 (2000).
56. Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.* **12**, 46–53 (2005).

57. Privalov, P. L., Dragan, A. I. & Crane-Robinson, C. The cost of DNA bending. *Trends Biochem. Sci.* **34**, 464–470 (2009).
58. Ye, Z. et al. Genome-wide analysis reveals positional-nucleosome-oriented binding pattern of pioneer factor FOXA1. *Nucleic Acids Res.* **44**, 7540–7554 (2016).
59. Cirillo, L. A. et al. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9**, 279–289 (2002).

Very-high-energy particle acceleration powered by the jets of the microquasar SS 433

A. U. Abeysekara¹, A. Albert², R. Alfaro³, C. Alvarez⁴, J. D. Álvarez⁵, R. Arceo⁴, J. C. Arteaga-Velázquez⁵, D. Avila Rojas³, H. A. Ayala Solares⁶, E. Belmont-Moreno³, S. Y. BenZvi^{7,35}, C. Brisbois⁸, K. S. Caballero-Mora⁴, T. Capistrán⁹, A. Carramiñana⁹, S. Casanova¹⁰, M. Castillo⁵, U. Cotti⁵, J. Cotzomi¹¹, S. Coutiño de León⁹, C. De León¹², E. De la Fuente¹², J. C. Díaz-Vélez^{13,14}, S. Dichiaro¹³, B. L. Dingus^{2,35}, M. A. DuVernois¹⁴, R. W. Ellsworth¹⁵, K. Engel¹⁶, C. Espinoza³, K. Fang^{17,18,35}, H. Fleischhack⁸, N. Fraija¹³, A. Galván-Gómez¹³, J. A. García-González³, F. Garfias¹³, A. González-Muñoz³, M. M. González¹³, J. A. Goodman¹⁶, Z. Hampel-Arias^{14,19}, J. P. Harding², S. Hernandez³, J. Hinton²⁰, B. Hona⁸, F. Hueyotl-Zahuanitla⁴, C. M. Hui²¹, P. Hüntemeyer⁸, A. Iriarte¹³, A. Jardin-Blicq²⁰, V. Joshi²⁰, S. Kaufmann⁴, P. Kar¹, G. J. Kunde², R. J. Lauer²², W. H. Lee¹³, H. León Vargas³, H. Li², J. T. Linnemann²², A. L. Longinotti⁹, G. Luis-Raya²³, R. López-Coto²⁴, K. Malone⁶, S. S. Marinelli²², O. Martinez¹¹, I. Martinez-Castellanos¹⁶, J. Martínez-Castro²⁵, J. A. Matthews²⁶, P. Miranda-Romagnoli²⁷, E. Moreno¹¹, M. Mostafá⁶, A. Nayerhoda¹⁰, L. Nellen²⁸, M. Newbold¹, M. U. Nisa⁷, R. Noriega-Papaqui²⁷, J. Pretz⁶, E. G. Pérez-Pérez²³, Z. Ren²⁶, C. D. Rho^{7,35*}, C. Rivière¹⁶, D. Rosa-González⁹, M. Rosenberg⁶, E. Ruiz-Velasco²⁰, F. Salesa Greus¹⁰, A. Sandoval³, M. Schneider²⁹, H. Schoorlemmer²⁰, M. Seglar Arroyo⁶, G. Sinnis², A. J. Smith¹⁶, R. W. Springer¹, P. Surajbali²⁰, I. Taboada³⁰, O. Tibolla⁴, K. Tollefson²², I. Torres⁹, G. Vianello³¹, L. Villaseñor¹¹, T. Weisgarber¹⁴, F. Werner²⁰, S. Westerhoff¹⁴, J. Wood¹⁴, T. Yapici⁷, G. Yodh³², A. Zepeda³³, H. Zhang^{34,35} & H. Zhou^{2,35*}

SS 433 is a binary system containing a supergiant star that is overflowing its Roche lobe with matter accreting onto a compact object (either a black hole or neutron star)^{1–3}. Two jets of ionized matter with a bulk velocity of approximately $0.26c$ (where c is the speed of light in vacuum) extend from the binary, perpendicular to the line of sight, and terminate inside W50, a supernova remnant that is being distorted by the jets^{2,4–8}. SS 433 differs from other microquasars (small-scale versions of quasars that are present within our own Galaxy) in that the accretion is believed to be super-Eddington^{9–11}, and the luminosity of the system is about 10^{40} ergs per second^{2,9,12,13}. The lobes of W50 in which the jets terminate, about 40 parsecs from the central source, are expected to accelerate charged particles, and indeed radio and X-ray emission consistent with electron synchrotron emission in a magnetic field have been observed^{14–16}. At higher energies (greater than 100 giga-electronvolts), the particle fluxes of γ -rays from X-ray hotspots around SS 433 have been reported as flux upper limits^{6,17–20}. In this energy regime, it has been unclear whether the emission is dominated by electrons that are interacting with photons from the cosmic microwave background through inverse-Compton scattering or by protons that are interacting with the ambient gas. Here we report teraelectronvolt γ -ray observations of the SS 433/W50 system that spatially resolve the lobes. The teraelectronvolt emission is localized to structures in the lobes, far from the centre of the system where the jets are formed. We have measured photon energies of at least 25 teraelectronvolts, and these are certainly not

Doppler-boosted, because of the viewing geometry. We conclude that the emission—from radio to teraelectronvolt energies—is consistent with a single population of electrons with energies extending to at least hundreds of teraelectronvolts in a magnetic field of about 16 microgauss.

In the SS 433/W50 complex, several regions located west of the central binary (w1 and w2) and east (e1, e2, e3) are observed to emit hard X-rays⁶. Previous searches for very high-energy (VHE) γ -ray emission from the hotspots between roughly 100 GeV and 10 TeV have produced null results^{17–20}, though an excess observed at about 800 MeV may be associated with SS 433 and W50²¹. The High Altitude Water Cherenkov (HAWC) observatory, Mexico, is a wide field-of-view VHE γ -ray observatory surveying the Northern Hemisphere above 1 TeV, and is optimized for photon detection above 10 TeV²². SS 433 transits 15° from the zenith of the HAWC detector each day, and has been observed with >90% uptime since the start of detector operations in 2015.

In 1,017 days of measurements with HAWC, an excess of γ -rays with a post-trials significance of 5.4σ has been observed in a joint fit of the eastern and western interaction regions of the jets of SS 433. The emission is plotted in galactic coordinates in Fig. 1, which includes an overlay of the X-ray observations of the jets and the central binary. The γ -ray emission is spatially coincident with the X-ray hotspots w1 and e1; no significant emission is observed at the location of the central binary where the jets are produced.

Spatial and spectral fits to SS 433 are performed in a semicircular region of interest (ROI) designed to mask out diffuse emission from

¹Department of Physics and Astronomy, University of Utah, Salt Lake City, UT, USA. ²Physics and Theoretical Divisions, Los Alamos National Laboratory, Los Alamos, NM, USA. ³Instituto de Física, Universidad Nacional Autónoma de México, Mexico City, Mexico. ⁴Universidad Autónoma de Chiapas, Tuxtla Gutiérrez, Mexico. ⁵Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico. ⁶Department of Physics, Pennsylvania State University, University Park, PA, USA. ⁷Department of Physics and Astronomy, University of Rochester, Rochester, NY, USA. ⁸Department of Physics, Michigan Technological University, Houghton, MI, USA. ⁹Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico. ¹⁰Institute of Nuclear Physics Polish Academy of Sciences, IFJ-PAN, Krakow, Poland. ¹¹Max-Planck Institute for Nuclear Physics, Heidelberg, Germany. ¹²Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico. ¹³Departamento de Física, Centro Universitario de Ciencias Exactas e Ingenierías, Universidad de Guadalajara, Guadalajara, Mexico. ¹⁴Department of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin-Madison, Madison, WI, USA. ¹⁵Instituto de Astronomía, Universidad Nacional Autónoma de México, Mexico City, Mexico. ¹⁶School of Physics, Astronomy, and Computational Sciences, George Mason University, Fairfax, VA, USA. ¹⁷Department of Physics, University of Maryland, College Park, MD, USA. ¹⁸Department of Astronomy, University of Maryland, College Park, MD, USA. ¹⁹Joint Space-Science Institute, University of Maryland, College Park, MD, USA. ²⁰Inter-university Institute for High Energies, Université Libre de Bruxelles, Brussels, Belgium. ²¹NASA Marshall Space Flight Center, Astrophysics Office, Huntsville, AL, USA. ²²Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, USA. ²³Department of Physics and Astronomy, Michigan State University, East Lansing, MI, USA. ²⁴Universidad Politécnica de Pachuca, Pachuca, Mexico. ²⁵INFN and Università di Padova, Padova, Italy. ²⁶Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico. ²⁷Universidad Autónoma del Estado de Hidalgo, Pachuca, Mexico. ²⁸Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico. ²⁹Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, CA, USA. ³⁰School of Physics and Center for Relativistic Astrophysics, Georgia Institute of Technology, Atlanta, GA, USA. ³¹Department of Physics, Stanford University, Stanford, CA, USA. ³²Department of Physics and Astronomy, University of California, Irvine, Irvine, CA, USA. ³³Physics Department, Centro de Investigación y de Estudios Avanzados del IPN, Mexico City, Mexico. ³⁴Department of Physics and Astronomy, Purdue University, West Lafayette, IN, USA. ³⁵These authors contributed equally: S. Y. BenZvi, B. L. Dingus, K. Fang, C. D. Rho, H. Zhang, H. Zhou. *e-mail: crho2@ur.rochester.edu; hao@lanl.gov

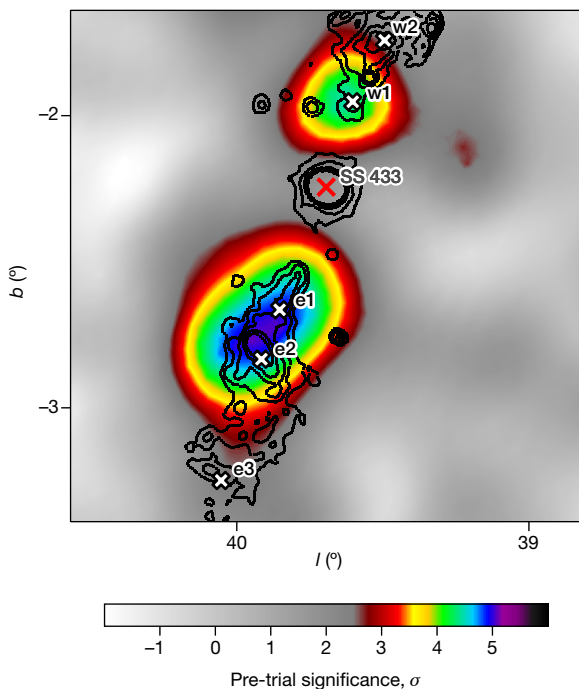


Fig. 1 | VHE γ -ray image of the SS 433/W50 region in Galactic coordinates. The colour scale indicates the statistical significance of the excess counts above the background of nearly isotropic cosmic rays before accounting for statistical trials. The figure shows the γ -ray excess measured after the fitting and subtraction of γ -rays from the spatially extended source MGRO J1908+06. The jet termination regions e1, e2, e3, w1 and w2 observed in the X-ray data are indicated, as well as the location of the central binary. The solid contours show the X-ray emission observed from this system.

the Galactic plane. The ROI also removes significant spatially extended emission from the nearby γ -ray source MGRO J1908+06. The spatial distribution and spectrum of γ -rays from MGRO J1908+06 are fitted using an electron diffusion model²³, and point-like sources centred on e1 and w1 are fitted on top of this extended emission. As a systematic check, the regions are also fitted using X-ray spatial templates and extended Gaussian functions. Neither improves the statistical significance of the fits. Upper limits on the angular size of the emission regions are 0.25° for the east hotspot and 0.35° for the west hotspot at 90% confidence. Given the distance to the source of 5.5 kpc, this corresponds to a physical size of 24 pc and 34 pc, respectively. The constraint is tighter on the eastern hotspot owing to its higher statistical significance.

The VHE γ -ray flux is consistent with a hard E^{-2} spectrum, though current data from HAWC are not of sufficient significance to constrain the spectral index. Therefore, we report the flux of both hotspots at 20 TeV, at which systematic uncertainties due to the choice of spectral model are minimized and the sensitivity of HAWC is maximized. At e1, the VHE flux is $2.4^{+0.6}_{-0.5}(\text{stat.})^{+1.3}_{-1.3}(\text{syst.}) \times 10^{-16} \text{ TeV}^{-1} \text{ cm}^{-2} \text{ s}^{-1}$, and at w1 the flux is $2.1^{+0.6}_{-0.5}(\text{stat.})^{+1.2}_{-1.2}(\text{syst.}) \times 10^{-16} \text{ TeV}^{-1} \text{ cm}^{-2} \text{ s}^{-1}$. HAWC detects γ -rays from the interaction regions up to at least 25 TeV. The energies of these γ -rays are a factor of three to ten higher than previous measurements from microquasars^{24,25}. Since most γ -ray telescopes are optimized for measurements below 10 TeV, this may explain why these photons were not observed in previous observational campaigns.

The γ -rays detected by HAWC are produced by radiative or decay processes from particles of much higher energy. The detection yields important information about the mechanisms and sites of particle acceleration, the types of particles accelerated (for example, protons or electrons), and the radiative processes that produce the spectrum of emission from radio to VHE γ -rays. Two scenarios for explaining the

HAWC observations of the e1 and w1 regions can be tested. The first is that protons are primarily responsible for the observed γ -rays. Protons must have an energy of at least 250 TeV to produce 25-TeV γ -rays through hadronic collisions with ambient gas. Proton–proton collisions yield neutral pions (π^0) that decay to VHE γ -rays, and charged pions (π^\pm) that decay to the secondary electrons and positrons responsible for radio to X-ray emission via synchrotron radiation. This scenario is of particular interest because there is spectroscopic evidence for ionized nuclei in the inner jets of SS 433^{8,26}. The alternative scenario requires electrons of at least 130 TeV to up-scatter the low-energy photons from the cosmic microwave background (CMB) to 25-TeV γ rays. In this case, the radio to X-ray emission is dominated by synchrotron radiation from the same population of electrons in the magnetized plasma of the jets and lobes.

The fact that the VHE emission is detected along a line of sight nearly orthogonal to the jet axis means that charged particle trajectories become isotropic before they interact to produce the γ -rays. The embedded magnetic fields in the VHE regions can easily deflect the accelerated particles because their typical gyroradii are much smaller than the size of the emission regions, approximately 30 pc. The jets are only mildly relativistic, so the emission from the interaction regions will have a negligible Doppler beaming effect and remain nearly isotropic.

The flux of VHE γ -rays observed by HAWC makes the proton scenario for SS 433 unlikely, because the total energy required to produce the highly relativistic protons is too high. The jets of SS 433 are known to be radiatively inefficient, with most of the jet energy transformed into the thermal energy of W50^{16,27} rather than into particle acceleration. We model the primary proton spectrum as a power law with an exponential cutoff, $dN/dE_p \propto E_p^{-2} \exp(-E_p/1 \text{ PeV})$. If we assume that 10% of the jet kinetic energy converts into accelerated protons, and that the ambient gas density^{16,27} is 0.05 cm^{-3} , then the resulting flux of γ -rays from proton–proton collisions is much less than the observed γ -ray flux, as shown in the dash-dotted line of Fig. 2. In fact, for a target proton density as large as 0.1 cm^{-3} in the e1 region^{16,27}, the total energy of the proton population needs to be around $3 \times 10^{50} \text{ erg}$ to explain the observed γ -rays, assuming an E_γ^{-2} spectrum. This is comparable to the total jet energy available during the presumed 30,000-year lifetime² of SS 433. Furthermore, because the synchrotron emission from secondary electrons from charged pion decay is always lower than the γ -ray flux from π^0 decay, and the observed X-ray flux is higher than the γ -ray flux, the X-rays cannot originate solely from secondary electrons. Finally, the proton scenario requires that the protons remain trapped in the region observed by HAWC for the lifetime² of SS 433. This means the protons must diffuse very slowly, with a diffusion coefficient of about $1/1,000$ of the typical value²⁸ of the interstellar medium (ISM), $D_{\text{ISM}} \approx 3 \times 10^{28} (E/3 \text{ GeV})^{1/3} \text{ cm}^2 \text{ s}^{-1}$. This value, comparable to the theoretical Bohm limit, is very small but not impossible. Given the uncertainties in the historical jet flux, the ambient particle density and the radiative efficiency, we cannot exclude the possibility that some fraction of the γ -ray flux is produced by protons. However, we do rule out the possibility that the VHE γ -rays are entirely produced by protons.

Highly relativistic electrons, on the other hand, can produce γ -rays much more efficiently, primarily via inverse Compton scattering of CMB photons to γ -rays. The inverse Compton losses due to upscattering of infrared and optical photons are suppressed owing to the Klein–Nishina effect and are thus dominated by scattering of CMB photons²⁹. In this scenario, the ratio of the VHE γ -ray to X-ray fluxes constrains the energy density in the magnetic field compared to the energy density in CMB photons. We have modelled the broadband spectral energy distribution of the eastern emission region $15'$ to $33'$ from the centre of SS 433. The solid and dashed lines in Fig. 2 show the spectral energy distribution of a leptonic model for e1 produced by an injected flux of relativistic electrons with an energy spectrum $dN/dE \propto E^{-\alpha} \exp(-E/E_{\text{max}})$ in a magnetic field of strength B . We use the parameters $\alpha = 1.9$, $E_{\text{max}} = 3.5 \text{ PeV}$, and $B = 16 \mu\text{G}$ (see Methods). The estimate of the magnetic field strength is consistent with the

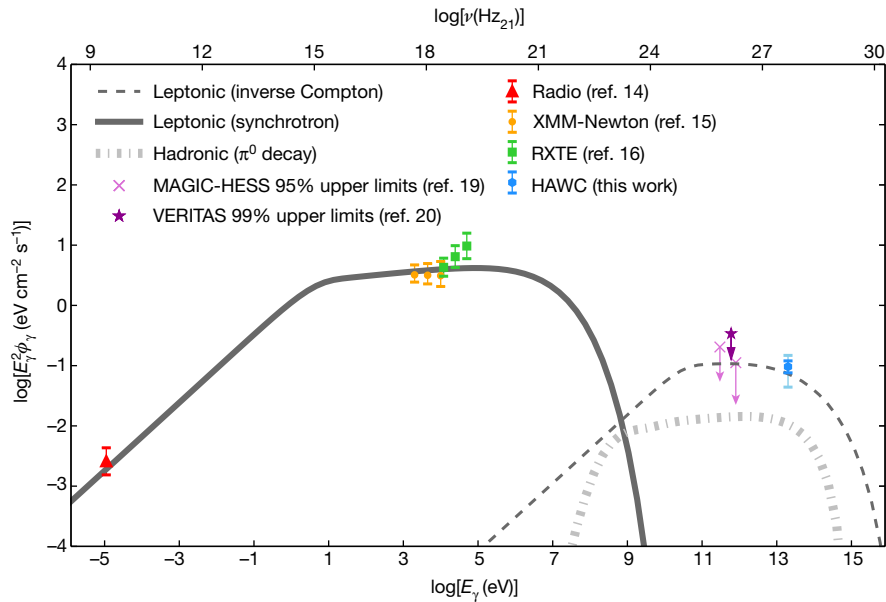


Fig. 2 | Broadband spectral energy distribution of the eastern emission region e1. The data include radio¹⁴, soft X-ray¹⁵ and VHE γ -ray upper limits^{19,20}, and HAWC observations of e1. Error bars indicate 1σ uncertainties, with the thick (thin) errors on the HAWC flux indicating statistical (systematic) uncertainties and arrows indicating flux upper limits. The multiwavelength spectrum produced by electrons assumes a single electron population following a power-law spectrum

equipartition of energy between the relativistic electrons and magnetic fields, which is common in astrophysical systems¹⁶. The required total energy budget for relativistic electrons is three orders of magnitude lower than the total jet energy.

The maximum electron energy of about 1 PeV has important implications for electron acceleration sites and acceleration mechanisms in SS 433. SS 433 is distinguished from other binary systems with relativistic objects because it achieves a supercritical accretion of gas onto the central engine (the compact object)². Powerful accretion flows and the inner jets near the compact object have therefore been proposed as possible acceleration sites of relativistic particles²⁶. However, the observation from HAWC suggests that ultrarelativistic electrons are not accelerated near the centre of the binary. If the electrons were accelerated in the central region, they would have cooled by the time they reached the sites of observed VHE emission. Owing to their small gyroradii, high-energy electrons may transport in a magnetized medium via diffusion or advection. The distance travelled via diffusion within the cooling time t_{cool} of an electron of energy E moving in a magnetic field of strength B is $r_d = 2\sqrt{Dt_{\text{cool}}} \approx 36 \text{ pc } (E/1 \text{ PeV})^{-1/3} (B/16 \mu\text{G})^{-1}$, using the diffusion coefficient D typical of the ISM²⁸. This distance would be even smaller for diffusion coefficients lower than the ISM value. Similarly, the distance travelled by electrons being advected with the jet flow is $r_{\text{adv}} = 0.26c \times t_{\text{cool}} \approx 4 \text{ pc } (E/1 \text{ PeV})^{-1} (B/16 \mu\text{G})^{-2}$ for a jet velocity of $0.26c$. Both distance scales are smaller than the 40-pc distance between the binary and e1, indicating that the electrons are not accelerated near the centre of the system.

Instead, the highly energetic electrons in SS 433 are probably accelerated in the jets and near the VHE γ -ray emission regions. This presents a challenge to current acceleration models. For example, particle acceleration may be driven by the dissipation of the magnetic fields in the jets, but above several hundred teraelectronvolts the electron acceleration time exceeds the electron cooling time, assuming a $16\text{-}\mu\text{G}$ magnetic field. Thus, the system does not appear to have sufficient acceleration power, unless there are very concentrated magnetic fields along the jets. If instead particle acceleration is driven by standing shocks produced by the bulk flow of the jets, it is possible to reach

with an exponential cutoff. The electrons produce radio to X-ray photons through synchrotron emission in a magnetic field (thick solid line) and teraelectronvolt γ rays through inverse Compton scattering of the CMB (thin dashed line). The dash-dotted line represents the radiation produced by protons, assuming that 10% of the jet kinetic energy converts into protons.

petaelectronvolt energies if the size of the acceleration region is larger than the gyroradii of the electrons. However, shocks in the interaction regions are not currently resolved by X-ray or γ -ray measurements.

Studies of microquasars such as SS 433 provide valuable probes of the particle acceleration mechanisms in jets, since these objects are believed to be scale models of the much larger and more powerful jets in active galactic nuclei³⁰. Active galactic nuclei are the most prevalent VHE extragalactic sources and are believed to be the sources of the highest-energy cosmic rays. Although active galactic nuclei are not spatially resolved at VHE energies, with this observation we have identified a VHE source in which we can image the particle acceleration powered by jets. Future high-resolution observations of SS 433 are possible using atmospheric Cherenkov telescopes pointed to localize the emission sites better, and further high-energy measurements with HAWC will record the spectrum at high energies and better constrain the maximum energy of accelerated particles.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0565-5>.

Received: 24 May 2018; Accepted: 10 August 2018;
Published online 3 October 2018.

1. Margon, B. Observations of SS 433. *Annu. Rev. Astron. Astrophys.* **22**, 507–536 (1984).
2. Fabrika, S. The jets and supercritical accretion disk in SS433. *Astrophys. Space Phys. Rev.* **12**, 1–152 (2004).
3. Cherepashchuk, A. M. et al. INTEGRAL observations of SS433: results of coordinated campaign. *Astron. Astrophys.* **437**, 561–573 (2005).
4. Zealey, W. J., Dopita, M. A. & Malin, D. F. The interaction between the relativistic jets of SS433 and the interstellar medium. *Mon. Not. R. Astron. Soc.* **192**, 731–743 (1980).
5. Margon, B. & Anderson, S. F. Ten years of SS 433 kinematics. *Astrophys. J.* **347**, 448–454 (1989).
6. Safi-Harb, S. & Ögelman, H. ROSAT and ASCA observations of W50 associated with the peculiar source SS 433. *Astrophys. J.* **483**, 868–881 (1997).
7. Eikenberry, S. S. et al. Twenty years of timing SS 433. *Astrophys. J.* **561**, 1027 (2001).
8. Migliari, S., Fender, R. P. & Mendez, M. Iron emission lines from extended X-ray jets in SS 433: reheating of atomic nuclei. *Science* **297**, 1673 (2002).

9. Mirabel, I. & Rodríguez, L. F. Sources of relativistic jets in the galaxy. *Annu. Rev. Astron. Astrophys.* **37**, 409–443 (1999).
10. Begelman, M. C., King, A. R. & Pringle, J. E. The nature of SS433 and the ultraluminous X-ray sources. *Mon. Not. R. Astron. Soc.* **370**, 399–404 (2006).
11. Fabrika, S., Ueda, Y., Vinokurov, A., Sholukhova, O. & Shidatsu, M. Supercritical accretion discs in ultraluminous X-ray sources and SS 433. *Nat. Phys.* **11**, 551 (2015).
12. Cherepashchuk, A. M., Aslanov, A. A. & Kornilov, V. G. WBVR photometry of SS 433—spectra of the normal star and the accretion disk. *Sov. Astron.* **26**, 697–702 (1982).
13. Tetarenko, B. E., Sivakoff, G. R., Heinke, C. O. & Gladstone, J. C. WATCHDOG: a comprehensive all-sky database of galactic black hole X-ray binaries. *Astrophys. J. Suppl.* **222**, 15 (2016).
14. Geldzahler, B. J., Pauls, T. & Salter, C. J. Continuum observations of the supernova remnants W50 and G 74.9+1.2 at 2695 MHz. *Astron. Astrophys.* **84**, 237–244 (1980).
15. Brinkmann, W., Pratt, G. W., Rohr, S., Kawai, N. & Burwitz, V. XMM-Newton observations of the eastern jet of SS433. *Astron. Astrophys.* **463**, 611–619 (2007).
16. Safi-Harb, S. & Petre, R. Rossi X-ray timing explorer observations of the eastern lobe of W50 associated with SS 433. *Astrophys. J.* **512**, 784–792 (1999).
17. Aharonian, F. et al. TeV gamma-ray observations of SS-433 and a survey of the surrounding field with the HEGRA IACT-System. *Astron. Astrophys.* **439**, 635–643 (2005).
18. Hayashi, S. et al. Search for VHE gamma rays from SS433/W50 with the CANGAROO-II telescope. *Astropart. Phys.* **32**, 112–119 (2009).
19. Ahnen, M. L. et al. Constraints on particle acceleration in SS433/W50 from MAGIC and H.E.S.S. observations. *Astron. Astrophys.* **612**, A14 (2018).
20. Kar, P. VERITAS observations of high-mass X-ray binary SS 433. *Proc. Sci.* (35th Int. Cosmic Ray Conf.) ICRC2017, <https://doi.org/10.22323/1.301.0713> (2018).
21. Bordas, P., Yang, R., Kafexhiu, E. & Aharonian, F. Detection of persistent gamma-ray emission toward SS433/W50. *Astrophys. J.* **807**, L8 (2015).
22. Abeysekara, A. U. et al. The 2HWC HAWC Observatory Gamma Ray Catalog. *Astrophys. J.* **843**, 40 (2017).
23. López-Coto, R. et al. Effect of the diffusion parameters on the observed γ -ray spectrum of sources and their contribution to the local all-electron spectrum: the EDGE code. *Astropart. Phys.* **102**, 1–11 (2018).
24. Albert, J. et al. Variable very high energy gamma-ray emission from the microquasar LS I +61° 303. *Science* **312**, 1771–1773 (2006).
25. Archambault, S. et al. Exceptionally bright TeV flares from the binary LS I +61° 303. *Astrophys. J.* **817**, L7 (2016).
26. Reynoso, M. M., Romero, G. E. & Christiansen, H. R. Production of gamma rays and neutrinos in the dark jets of the microquasar SS433. *Mon. Not. R. Astron. Soc.* **387**, 1745–1754 (2008).
27. Panferov, A. A. Jets of SS 433 on scales of dozens of parsecs. *Astron. Astrophys.* **599**, A77 (2017).
28. Ptuskin, V. S., Moskalenko, I. V., Jones, F. C., Strong, A. W. & Zirakashvili, V. N. Dissipation of magnetohydrodynamic waves on energetic particles: impact on interstellar turbulence and cosmic ray transport. *Astrophys. J.* **642**, 902–916 (2006).
29. Moderski, R., Sikora, M., Coppi, P. S. & Aharonian, F. A. Klein-Nishina effects in the spectra of non-thermal sources immersed in external radiation fields. *Mon. Not. R. Astron. Soc.* **364**, 1488 (2005).
30. Romero, G., Boettcher, M., Markoff, S. & Tavecchio, F. Relativistic jets in active galactic nuclei and microquasars. *Space Sci. Rev.* **207**, 5–61 (2017).

Acknowledgements We acknowledge support from: the US National Science Foundation (NSF); the US Department of Energy Office of High-Energy Physics; the Laboratory Directed Research and Development programme of Los Alamos National Laboratory; Consejo Nacional de Ciencia y Tecnología, México (grants 271051, 232656, 260378, 179588, 239762, 254964, 271737, 258865, 243290, 132197 and 281653) (Cátedras 873, 1563); Laboratorio Nacional HAWC de rayos gamma; L'OREAL Fellowship for Women in Science 2014; Red HAWC, México; DGAPA-UNAM (Dirección General Asuntos del Personal Académico—Universidad Nacional Autónoma de México; grants IG100317, IN111315, IN111716-3, IA102715, IN109916 and IA102917); VIEP-BUAP (Vicerrectoría de Investigación y Estudios de Posgrado-Benemérita Universidad Autónoma de Puebla); PIFI (Programa Integral de Fortalecimiento Institucional) 2012 and 2013; PRO-FOCIE (Programa de Fortalecimiento de la Calidad en Instituciones Educativas) 2014 and 2015; the University of Wisconsin Alumni Research Foundation; the Institute of Geophysics, Planetary Physics, and Signatures at Los Alamos National Laboratory; Polish Science Centre grant DEC-2014/13/B/ST9/945 and DEC-2017/27/B/ST9/02272; and Coordinación de la Investigación Científica de la Universidad Michoacana. We thank S. Delay, L. Díaz and E. Murrieta for technical support. We thank R. Mushotzky for providing the spectrum of the XMM-Newton data in the HAWC detection region.

Reviewer information Nature thanks A. Achterberg and M. Bowler for their contribution to the peer review of this work.

Author contributions C.D.R. and H. Zhou analysed the data and performed the maximum likelihood analysis. Multiwavelength modelling of the leptonic and hadronic emission was carried out by K.F. and H. Zhang. S.Y.B. and B.L.D. helped to prepare the manuscript. The entire HAWC Collaboration contributed through the construction, calibration and operation of the detector, the development and maintenance of reconstruction and analysis software, and the vetting of the analysis presented in this manuscript. All authors reviewed, discussed and commented on the results and the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0565-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.D.R. and H. Zhou.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reduction and maximum likelihood analysis. This analysis uses 1,017 days of data from the High Altitude Water Cherenkov (HAWC) Observatory collected between 26 November 2014 and 20 December 2017. The HAWC Observatory is an array of 300 tightly packed identical water Cherenkov detectors deployed 4,100 m above sea level on the slope of the volcano Sierra Negra, Mexico³¹. Each detector is a cylindrical water tank standing 5 m tall and 7.3 m in diameter, filled with 180,000 litres of purified water. At the bottom of each tank are four photomultipliers (PMTs) facing upward. The PMTs record the Cherenkov photons created by the relativistic secondary particles produced when primary cosmic rays and γ -rays interact at the top of the atmosphere. The HAWC array covers 22,000 m². Its construction ended in December 2014, and the full array was commissioned in March 2015.

Using the relative arrival time of photoelectrons (hits) detected by the PMTs, the arrival direction of primary γ -rays can be reconstructed with an accuracy³² of around 1° below 1 TeV to <0.2° above 10 TeV. The accuracy of the reconstruction determines the point spread function of the detector, and is a function of the energy, zenith angle and composition of the primary particle. Air showers from γ -rays are discriminated from the nearly isotropic background of hadronic cosmic rays by filtering out ‘clumpy’ patterns of hits, which are characteristic of the energy deposited in hadronic air showers. The cosmic-ray background rejection efficiency improves rapidly as a function of energy³², increasing from 90% at 1 TeV to 99.9% at 10 TeV.

To compute the statistical significance of γ -ray emission observed with HAWC, a maximum likelihood fit using parametric spatial and spectral models is applied to the data^{33,34}. The models are forward-folded through the detector response to produce expected counts of γ -ray signal events and cosmic-ray background events. The expectation is then compared to the observed counts N_{obs} . To calculate the expected counts as a function of position on the sky, the events are binned in a fine mesh using the HEALPix pixelization of the unit sphere³⁵. The pixelization is chosen to be 0.1°, roughly two to five times smaller than the radius of the instrument point spread function. To apply models of the energy spectrum of a source, the data are binned according to the fraction of PMTs in the detector triggered by an air shower³². This measure of shower ‘size’ is used as a coarse proxy for the energy of the primary particle; a total of nine size bins labelled $B = 1-9$ is used.

Given a model θ with spatial and spectral parameters, we maximize the likelihood of the model having produced the data as follows:

$$\ln \mathcal{L}(N_{\text{obs}}|\theta) = \sum_{B=1}^9 \sum_{j=1}^m \ln P(N_{\text{obs}}^{j,B}|\theta)$$

where the sum runs over the size bins B and the HEALPix pixels j in the region of interest (ROI) of the fit. P is the Poisson probability of detecting $N_{\text{obs}}^{j,B}$ events in pixel j and size bin B given the model parameters θ .

Within the ROI around SS 433 defined in Extended Data Fig. 1, two fits are performed to maximize the likelihood: a fit which accounts only for the emission from MGRO J1908+06 (null hypothesis), and a fit that accounts for the combined emission from MGRO J1908+06 and the SS 433 region (alternative hypothesis). The ratio of the maximum likelihood defines a test statistic (TS):

$$\text{TS} = 2(\ln \mathcal{L}(N_{\text{obs}}|\theta_{\text{alt}}) - \ln \mathcal{L}(N_{\text{obs}}|\theta_0))$$

where θ_0 and θ_{alt} represent the spatial and spectral parameters of the null and alternative hypotheses, respectively. TS is then converted to a P value to estimate the statistical significance of emission from SS 433. As discussed in the main text, the alternative hypothesis assumes two point sources with power-law spectra $dN/dE = f_0(E/20 \text{ TeV})^{-2}$, where the flux normalization f_0 is the free parameter of the spectral model.

Wilks’ Theorem is used to convert TS to a P value³⁶. In the joint likelihood maximization, there are two degrees of freedom (d.o.f.) for the two separately fitted flux normalizations of the hotspots at w1 and e1. Therefore, we calculate the one-tailed P value $\text{pr}(\text{TS} > \chi^2 = 41.2 | \text{d.o.f.} = 2) = 1.13 \times 10^{-9}$. This is the tail probability of $\text{TS} > 41.2$ assuming that, under the null hypothesis of no excess γ rays, TS follows a χ^2 distribution with 2 degrees of freedom. Since the positions of the point source fits at w1 and e1 were chosen after looking into the data, and because we are searching for other microquasars in the field of view of HAWC, we must apply a posteriori corrections to the P value to account for multiple-comparison effects.

The X-ray interaction regions w1, w2, e1, e2 and e3 are a priori candidates for the locations of the maxima, as is the centre of the binary system, for a total of six potential hotspots. Given the angular resolution of HAWC, it would not be possible to spatially resolve all six hotspots; at best three regions (east, west and centre) can be separately fitted with confidence. There are 23 possible combinations of the six a priori locations that can be used to fit one, two or three hotspots in the eastern, central and western regions of the source. We add an additional 12 trials

to account for the known microquasars in the field of view of HAWC^{9,13,37}. This trial factor is conservative given that several Galactic microquasars are already known teraelectronvolt sources^{38,39}. Given 35 total trials, the corrected P value is 3.96×10^{-8} , which corresponds to a statistical significance of 5.4σ .

Modelling of the nearby extended source MGRO J1908+06. A bright extended source, MGRO J1908+06, is detected with more than 30σ in this dataset and is located less than 2° from the γ -ray hotspots of SS 433 (Extended Data Fig. 1). The region of MGRO J1908+06 contains a pulsar and a supernova remnant, but it is not clear whether the observed teraelectronvolt γ -ray emission is from either or both of them. A detailed discussion of MGRO J1908+06 is beyond the scope of this paper. However, the morphology of MGRO J1908+06 must be carefully studied to minimize the contamination of the emission due to MGRO J1908+06 on the fluxes of the lobes.

A maximum likelihood analysis is performed that simultaneously fits the emission from MGRO J1908+06 and the hotspots at w1 and e1. An electron diffusion model appropriate for older pulsar wind nebulae^{23,40} is used to describe the spatial morphology of MGRO J1908+06. Given the uncertainty of the nature of MGRO J1908+06, two other spatial models with Gaussian and power law radial profiles are also tested in the simultaneous fit. The choice of spatial model affects the best-fit fluxes from e1 and w1 at the level of $\pm 20\%$. We adopt this value as a systematic uncertainty on the flux from w1 and e1 due to VHE emission from the nearby extended source.

Contamination from galactic diffuse emission. The e1 and w1 regions are located roughly 2° from the Galactic plane, so the contamination from the Galactic diffuse emission (GDE) is negligible. However, MGRO J1908+06 has a Galactic latitude of about 1° . Since the three spatial models used to fit MGRO J1908+06 are radially symmetric, the presence of GDE has the potential to produce an overestimate in the flux from MGRO J1908+06, which could result in an underestimate in the flux measured from w1 and e1. To minimize the effect of the GDE on the fit, the ROI is defined to be a semicircular region centred on the position of MGRO J1908+06 (Extended Data Fig. 1). The ROI is designed to reduce the effect of GDE by excluding the half of the source closest to the Galactic plane.

To estimate the systematic uncertainties associated with the choice of ROI and possible contamination from GDE, a second maximum likelihood fit is performed using a full-disk ROI that includes GDE, emission from MGRO J1908+06, and w1 and e1. The spatial distribution of the GDE is modelled with a Gaussian profile of three different widths of 0.5° , 1° and 2° in Galactic latitude, and is treated as constant in Galactic longitude over the width of the ROI. Comparing the results to the fit with a semicircular ROI indicates that contamination from GDE is less than 10% for e1 and less than 20% for w1.

Fit results. The fit results are reported in Extended Data Table 1. Fitting the emission from w1 and e1 simultaneously, we calculate $\text{TS} = 41.2$, which corresponds to a 5.4σ observation ($P = 3.96 \times 10^{-8}$) after accounting for a posteriori statistical trial factors. To check the consistency of the results, we also fit w1 and e1 separately, redefining the alternative model to include MGRO J1908+06 and only e1 or w1. The significance of the VHE excess from these locations is below 5σ in the fits, but the estimated fluxes are consistent with the simultaneous fit.

An additional check is made on the effect of fixing versus floating the best-fit positions of the emission from the east and west hotspots. In the original alternative hypothesis, the point sources are centred on e1 and w1. Here, the positions of the point sources are made additional free parameters in the point source fit. We find that allowing the positions of the teraelectronvolt hotspots to vary does not affect the flux estimates, which are consistent with the fixed-position fits. Moreover, the best-fit positions of the east and west teraelectronvolt emission regions are consistent with e1 and w1 within statistical uncertainties.

The choice of spectral model also affects the estimated γ -ray flux at 20 TeV. Extended Data Table 2 shows the dependence of the best-fit VHE flux from e1 and w1 on the assumed spectral models, including statistical uncertainties on the flux normalization at 20 TeV. Two spectral models were tested: a simple power law $dN/dE_\gamma \propto E_\gamma^{-\alpha}$, and a power law with an exponential cutoff $dN/dE_\gamma \propto E_\gamma^{-\alpha} \exp(-E_\gamma/E_{\text{cut}})$. The choice of spectral model can alter the flux normalization by almost a factor of two compared to the default E^{-2} model.

Summary of systematic uncertainties. The systematic uncertainties in the estimated fluxes from the teraelectronvolt hotspots of SS 433 include the following contributions: detector systematic effects, modelling ambiguities in MGRO J1908+06, and contamination from Galactic diffuse emission. The systematic uncertainties due to the modelling of MGRO J1908+06 and the contamination from Galactic diffuse emission are $\pm 20\%$ and -10% (-20%) for the east (west) hotspot, respectively, and are discussed in previous sections.

The detector response is estimated using Monte Carlo simulations and then optimized using observations of the Crab Nebula³², which appears point-like in the HAWC data. Systematic uncertainties that potentially affect the result presented here include the charge resolution and relative quantum efficiency of the PMTs, the absolute quantum efficiency of the PMTs, changes to the detector layout as

construction proceeded, uncertainties in the point spread function, and systematic differences in the distribution of arrival times of photoelectrons between data and simulation. The total systematic uncertainty on the flux normalization from detector effects is $\pm 50\%$.

All the components of the systematic uncertainties are summarized in Extended Data Table 3 and combined in quadrature to estimate the total systematic uncertainty on the VHE flux from w1 and e1. We note that since the systematic uncertainties due to MGRO J1908+06 and GDE are anti-correlated, the quadrature sum overestimates the total systematic uncertainty. However, the effect is not particularly important, since the detector systematic effects are the dominant source of uncertainty.

X-ray template fit and upper limit on the extent of the emission regions. We performed several maximum likelihood fits modelling the hotspots as spatially extended sources. In the first fit, we generated spatial templates for the eastern and western regions based on the X-ray contours published by ROSAT⁴¹ and then performed a joint likelihood fit with the two γ -ray hotspots and MGRO J1908+06. This produces no improvement in TS over a point-source fit.

To constrain the size of the γ -ray emission regions, likelihood fits are applied using a Gaussian morphology convolved with the point spread function of HAWC. To reduce the number of free parameters, we first fitted MGRO J1908+06 using an ROI with SS 433 and its hotspots excluded. The extended fit from MGRO J1908+06 is then subtracted from the data, and the residual γ -ray emission from the γ -ray hotspots is fitted using two Gaussian functions. The centres of the Gaussians are fixed to e1 and w1, and their angular widths are estimated in a simultaneous fit to both the eastern and western regions.

The maximum likelihood fit yields an angular width of 0.14 ± 0.06 degrees for the east hotspot and $0.08^{+0.14}_{-0.05}$ degrees for the west hotspot. We estimate the 90% confidence region on the extent as the value of Gaussian width that produces a decrease $\Delta TS = -2.71$ from the maximum likelihood value. The resulting 90% upper limits are 0.25° for the east region and 0.35° for the west region.

Upper limit on emission from the central binary. In the present dataset, no statistically significant emission is observed from the centre of SS 433. Using Feldman–Cousins likelihood ordering⁴², we estimate the 90% upper limit on the flux at 20 TeV to be $5.3 \times 10^{-17} \text{ TeV}^{-1} \text{ cm}^{-2} \text{ s}^{-1}$ after fitting MGRO J1908+06 and the emission at e1 and w1.

Upper limit on detected γ -ray energy. The binning of γ -ray events into size bins B causes us to lose information about the energies of the γ rays observed from SS 433. To determine the upper energy bound on the flux we observe, we scan over the maximum energy $E_{\gamma, \text{max}}$ used in the forward-folding analysis. Starting at $E_{\gamma, \text{max}} = 15 \text{ TeV}$, we find that TS increases monotonically until $E_{\gamma, \text{max}} = 25 \text{ TeV}$. Increasing $E_{\gamma, \text{max}}$ above this value causes TS to plateau (for e1) or decrease slightly (w1). We infer that the current measurement of e1 and w1 implies a minimum $E_{\gamma, \text{max}} = 25 \text{ TeV}$, and report this as a conservative estimate of the highest energy observed by HAWC.

Study of residual emission in the region of interest. As a final check of the quality of the maximum likelihood fits, we plot the distribution of the significance values in each HEALPix pixel in the ROI around SS 433 in Extended Data Fig. 2. The significance values are plotted in units of Gaussian σ . If only random background fluctuations are present, the significance values follow a standard normal distribution, shown by the dashed line in the figure.

Prior to the maximum likelihood fit, the significance distribution in the ROI is considerably skewed towards positive values owing to excess counts above background from MGRO J1908+06 and γ -rays from w1 and e1 (left panel of Extended Data Fig. 2). After the subtraction of the maximum likelihood fit to emission from MGRO J1908+06 (middle panel), the skew in the distribution is considerably reduced, though still visible owing to excess counts from the interaction regions near SS 433. Finally, subtraction of the maximum likelihood flux from w1 and e1 produces a distribution that, within statistical uncertainties, is equivalent to background fluctuations (right panel of Extended Data Fig. 2).

Multiwavelength modelling of the spectral energy distribution. To determine the origin of the VHE radiation, we model the broadband emission from e1 using 2.7 GHz radio measurements from the Effelsberg Telescope¹⁴, soft X-ray measurements between 2 keV and 10 keV from XMM-Newton¹⁵, hard X-ray measurements between 10 keV and 50 keV from the Rossi X-ray Timing Explorer¹⁶, VHE γ -ray upper limits from the MAGIC, H.E.S.S. and VERITAS imaging air Cherenkov telescopes^{19,20}, and HAWC observations. The data are taken from the region $15'$ to $33'$ east of SS 433, which covers the excess observed by HAWC centred at e1. Leptonic and hadronic models are used to produce spectral energy distributions for the low- and high-energy emission. The spectral energy distributions are shown in Fig. 2, and discussion of the two models is provided below.

VHE emission due to leptonic interactions. In the leptonic scenario considered in this paper, relativistic electrons scatter photons from the CMB photons to TeV energies via the inverse Compton process, and produce X-ray and radio emissions

by the synchrotron radiation. Although a far-infrared background in the lobe may contribute to the production of sub-TeV photons, no appreciable infrared emission has been reported near the γ -ray emission region^{2,43}. The electron spectrum is obtained by solving the continuity equation, considering radiative cooling⁴⁴. The best-fit values of the parameters of the injection spectrum, including the flux, spectral index and maximum energy of the electrons, and the magnetic field strength in the source region, are obtained through Markov Chain Monte Carlo⁴⁵ sampling of their likelihood distributions when fitting to the multiwavelength data. The radio and soft X-ray data points correspond to the Effelsberg Telescope¹⁴ and the XMM-Newton (the Mos1 detector)¹⁵ observations of a $6'$ circle centred on e1. A 30% uncertainty attributed to the unknown shape of the HAWC source is added to the statistical and systematic errors of the observational data, though we find that the uncertainty has a negligible impact on the fit. The hard X-ray data points and the sub-teraelectronvolt upper limits are set by the RXTE, MAGIC, H.E.S.S., and VERITAS observations of the e1 region^{16,19,20}.

The VHE flux is determined using the flux from e1 at 20 TeV reported in Extended Data Table 1, where separate fits were made to eastern and western hotspots, and the positions were fixed to e1 and w1. The best-fit values of the injection spectrum and magnetic field in the emission region are $\alpha = 1.87^{+0.04}_{-0.07}$, $\log[E_{\text{max}} (\text{PeV})] = 3.53^{+0.31}_{-0.38}$ and field strength $B = 16.04^{+2.60}_{-2.23} \mu\text{G}$. Taking the distance to the source to be 5.5 kpc, the fit suggests a total electron energy of $2.9 \times 10^{47} \text{ erg}$. This is a small fraction of the total energy deposited by the jets of SS 433 over their lifetime, which is about $9 \times 10^{50} \text{ erg}$ assuming a kinetic jet luminosity² of $10^{39} \text{ erg s}^{-1}$. Future multiwavelength observations dedicated to the VHE γ -ray emission region will better constrain the magnetic field strength and the properties of the electron population.

We note that the presence of several hundred teraelectronvolt-to-petaelectronvolt electrons would challenge the current particle acceleration mechanisms. Successful acceleration requires that the acceleration rate $\dot{\gamma} \approx eBv/m_e c^2$, where e is the charge of the electron and m_e is the electron mass, based on heuristic considerations (where v is the velocity associated with the notional electromotive force) exceed the cooling rate $\dot{\gamma} \approx 4\sigma c\tau_{\gamma}^{-1}(B^2/8\pi)/3m_e c^2$, assuming that synchrotron radiation dominates the cooling processes in the lobes of SS 433. This leads to a maximum electron energy $E_{e, \text{max}} = 271 \text{ TeV} (v/100 \text{ km s}^{-1})^{1/2} (B/16 \mu\text{G})^{-1/2}$. For reference, the Alfvén speed in the lobes is $v_A = 160 \text{ km s}^{-1} (n_b/0.05 \text{ cm}^{-3})^{-1/2} (B/16 \mu\text{G})$. A higher Alfvén speed could be achieved if the acceleration takes place in the central spine of the jet, where the mass loading due to black hole accretion is smaller and the magnetic field is stronger. Depending on the exact electron acceleration mechanisms, v could be associated with the jet flow velocity, or with the Alfvén speed. In both cases, using these estimates, it is possible that the maximum electron energy could exceed 1 PeV. However, the timescale of acceleration mechanisms such as second-order Fermi acceleration is proportional to $(v/c)^2$, making the production of several hundred teraelectronvolt electrons less efficient. Future VHE γ -ray and hard X-ray observations can better constrain the electron cutoff energy, and diagnose the in situ particle acceleration mechanism.

VHE emission due to hadronic interactions. In the hadronic scenario, high-energy protons interact with the ambient gas in the source, and produce γ rays via the decay $\pi^0 \rightarrow \gamma\gamma$. Extended Data Fig. 3 shows the fraction of jet power that needs to be converted to protons to produce the observed γ -ray flux. We assume a proton spectrum $dN/dE_p \propto E_p^{-\alpha} \exp(-E_p/E_{\text{max}})$, and adopt a proton–proton interaction cross-section⁴⁶ of around 50 mb, and a baryon density^{16,27} of $0.01\text{--}0.1 \text{ cm}^{-3}$. The total proton energy is obtained by integrating this spectrum normalized to the VHE γ -ray flux.

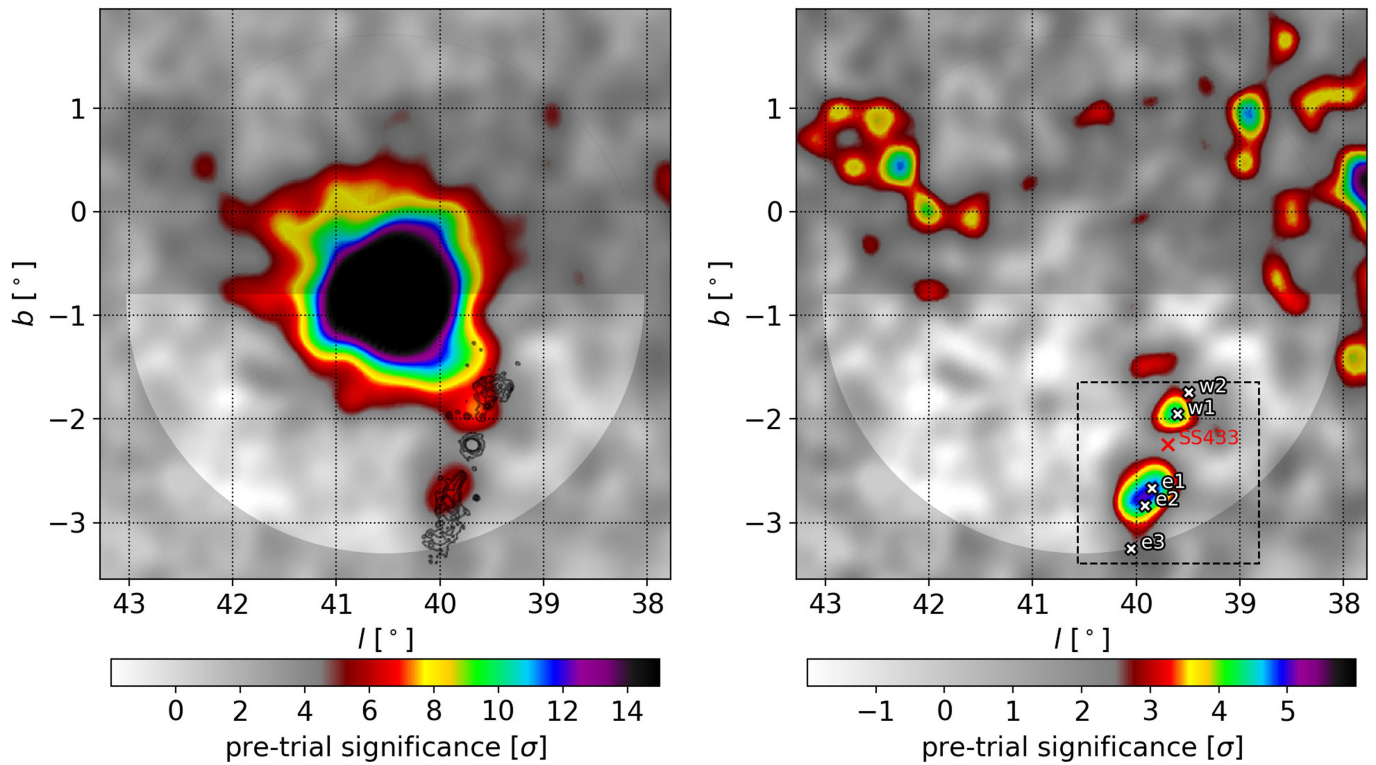
If the diffusion coefficient in the source is comparable to that in the ISM, no hadronic models would be allowed, because they would require a proton injection rate that exceeds the total kinetic luminosity of the jets of SS 433. Even in extreme circumstances, for example, where the diffusion coefficient is extremely small, possibly owing to scattering by turbulence generated from the streaming cosmic rays^{47,48}, particles could remain in the jet as long as the jet lifetime² of around 10^4 years. Assuming that protons follow a hard spectrum with $\alpha < 2$, the hadronic scenario would still require that at least 30% of the jet power goes to protons. Although a hadronic origin to the VHE flux is possible, it requires rather extreme source parameters and is therefore disfavoured.

Code availability. The study was carried out using the Analysis and Event Reconstruction Integrated Likelihood Fitting Framework (AERIE-LiFF) developed by the HAWC Collaboration. The software is open-source and publicly available on Github: <https://github.com/rjlauer/aerie-liff>. The code distribution includes instructions on installation and usage.

Data availability

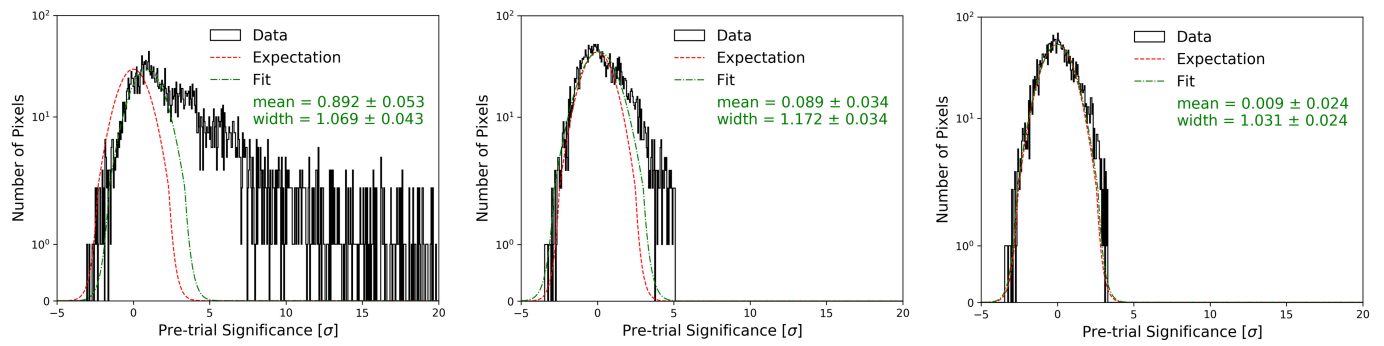
The datasets analysed during this study are available at a public repository maintained by the HAWC Collaboration: <https://data.hawc-observatory.org/>.

31. Smith, A. J. HAWC: design, operation, reconstruction and analysis. *Proc. Sci.* (34th Int. Cosmic Ray Conf.) ICRC2015, <https://doi.org/10.22323/1.236.0966> (2016).
32. Abeysekara, A. U. et al. Observation of the Crab Nebula with the HAWC Gamma Ray Observatory. *Astrophys. J.* **843**, 39 (2017).
33. Younk, P. W. et al. A high-level analysis framework for HAWC. *Proc. Sci.* (34th Int. Cosmic Ray Conf.) ICRC2015, <https://doi.org/10.22323/1.236.0948> (2016).
34. Vianello, G. et al. The multi-mission maximum likelihood framework. *Proc. Sci.* (34th Int. Cosmic Ray Conf.) ICRC2015, <https://doi.org/10.22323/1.236.1042> (2016).
35. Gorski, K. M. et al. HEALPix—a framework for high resolution discretization, and fast analysis of data distributed on the sphere. *Astrophys. J.* **622**, 759–771 (2005).
36. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).
37. Chaty, S. & Delautier, S. Microquasars. <http://www.aim.univ-paris7.fr/CHATY/Microquasars/microquasars.html> (Université Paris Diderot, 2006).
38. Aharonian, F. et al. Discovery of very high energy gamma rays associated with an X-ray binary. *Science* **309**, 746–749 (2005).
39. Aliu, E. et al. Multiwavelength observations of the TeV binary LS I +61 303 with VERITAS, Fermi-LAT, and Swift/XRT during a TeV outburst. *Astrophys. J.* **779**, 88 (2013).
40. Abeysekara, A. U. et al. Extended gamma-ray sources around pulsars constrain the origin of the positron flux at Earth. *Science* **358**, 911–914 (2017).
41. Brinkmann, W., Aschenbach, B. & Kawai, N. ROSAT observations of the W 50/SS 433 system. *Astron. Astrophys.* **312**, 306–316 (1996).
42. Feldman, G. J. & Cousins, R. D. A unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57**, 3873–3889 (1998).
43. Fuchs, Y., Mirabel, I. F. & Ogley, R. N. Mid-infrared observations of GRS 1915+105 and SS 433. *Astrophys. Space Sci. Suppl.* **276**, 99–100 (2001).
44. Finke, J. D. & Dermer, C. D. Cosmic ray electron evolution in the supernova remnant RX J1713.7-3946. *Astrophys. J.* **751**, 65 (2012).
45. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Publ. Astron. Soc. Pacif.* **125**, 306–312 (2013).
46. Particle Data Group. Review of particle physics. *Phys. Lett. B* **592**, 1–5 (2004).
47. Amato, E. & Blasi, P. Non linear particle acceleration at non-relativistic shock waves in the presence of self-generated turbulence. *Mon. Not. R. Astron. Soc.* **371**, 1251–1258 (2006).
48. Malkov, M. A., Diamond, P. H., Sagdeev, R. Z., Aharonian, F. A. & Moskalenko, I. V. Analytic solution for self-regulated collective escape of cosmic rays from their acceleration sites. *Astrophys. J.* **768**, 73 (2013).



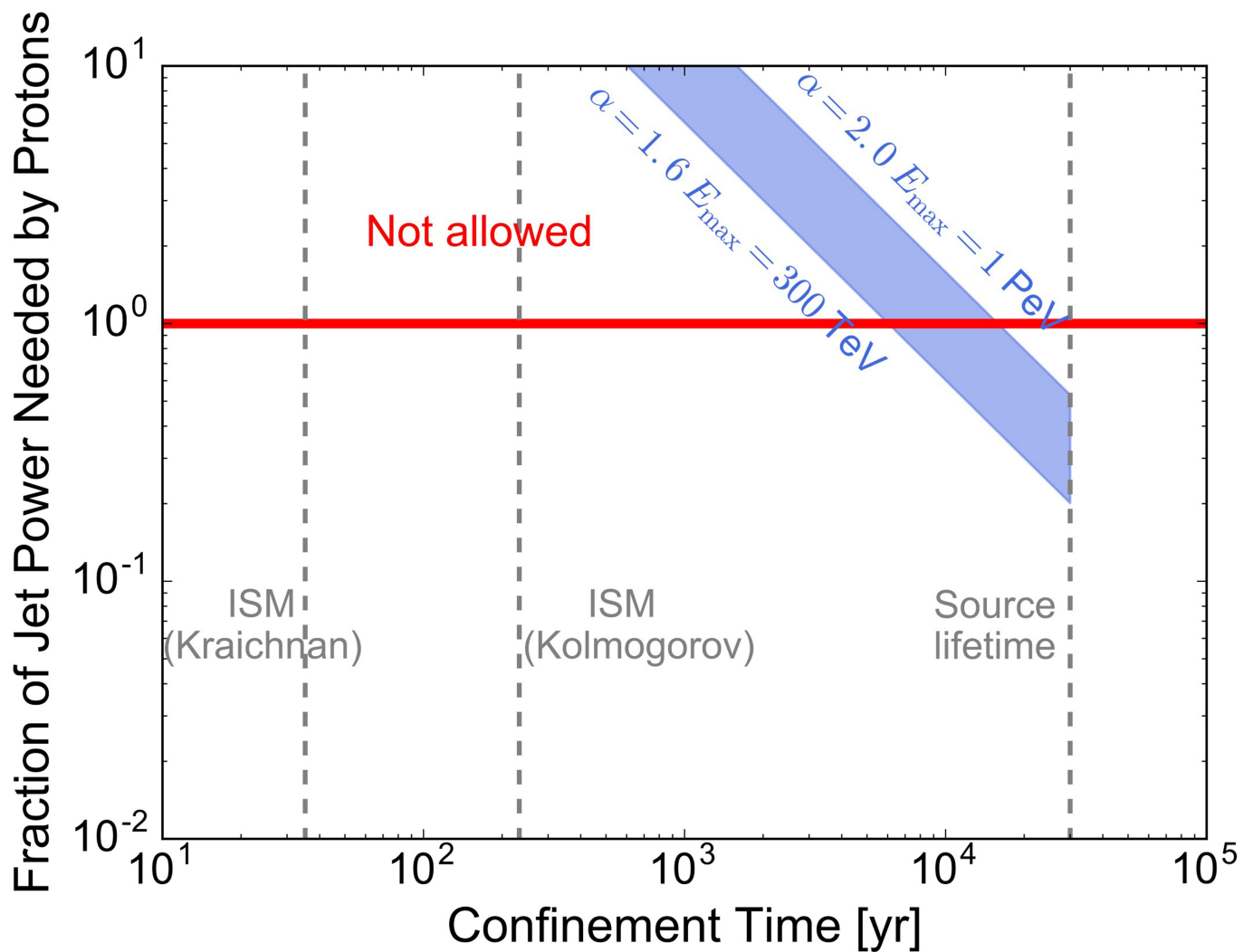
Extended Data Fig. 1 | VHE γ rays from MGRO J1908+06 and SS 433/W50. The colour scale indicates the statistical significance of the excess counts above the background of contaminating cosmic rays and γ -rays before accounting for statistical trials. The bright extended γ -ray source MGRO J1908+06 is shown at the centre of the left panel with SS 433/W50 at the bottom. The dark contours show X-ray emission from SS 433 and its

jets⁴¹. The semicircular area indicates the region of interest used to fit the γ -ray observations. The right panel shows the γ -ray excess measured after the fitting and subtraction of γ -rays from the spatially extended source MGRO J1908+06. The dashed box indicates the region shown in Fig. 1. The jet termination regions e1, e2, e3, w1 and w2 observed in the X-ray data are indicated, as well as the location of the central binary.



Extended Data Fig. 2 | Distribution of pixel significance in the region of interest of the fit. The significance is defined as deviations from the background expectation, in the HAWC sky map (left panel), after fitting

and subtraction of emission from MGRO J1908+06 (middle panel), and after fitting and removal of emission from MGRO J1908+06 and the γ rays from w1 and e1 (right panel).



Extended Data Fig. 3 | Fraction of jet power needed to produce the observed VHE γ -rays in the hadronic scenario. The blue-shaded region shows the energy injection rate of protons, in units of the kinetic luminosity of the jet, required to produce the observed VHE γ -rays by interacting with ambient gas, as a function of the proton confinement time. A gas density of 0.05 cm^{-3} is adopted for the source vicinity^{16,27}. Most hadronic models require $>100\%$ jet power (above the red solid line)

and are thus not allowed. Even when the diffusion coefficient is extremely small (for reference, the dashed grey lines show the source age and the confinement time of 200 TeV protons in a 30-pc region in the ISM with Kraichnan- and Kolmogorov-type diffusion) and when the spectral index is much harder than 2, the hadronic scenario still requires a large energy input from the jet.

Extended Data Table 1 | Fits to teraelectronvolt emission from SS 433 using nested point source models

Lobe	Position (RA, Dec)	dN/dE at 20 TeV [$10^{-16} \text{ TeV}^{-1} \text{ cm}^{-2} \text{ s}^{-1}$]	TS	Significance (post-trials)
Simultaneous fit to E+W hotspots, positions fixed.				
E	19:13:37 04°55'48"	$2.4^{+0.6+1.3}_{-0.5+1.3}$	41.2	5.4 σ
W	19:10:37 05°02'13"	$2.1^{+0.6+1.2}_{-0.5-1.2}$		
Separate fit to E+W hotspots, positions fixed.				
E	19:13:37 04°55'48"	$2.5^{+0.7+1.4}_{-0.5-1.4}$	24.3	4.6 σ
W	19:10:37 05°02'13"	$2.3^{+0.7+1.3}_{-0.5-1.3}$	20.4	4.2 σ
Separate fit to E+W hotspots, positions floated.				
E	19: 14: 11 $^{+20}_{-39}$ s 04°59'10" $^{+03/30}_{-06/18}$ "	$2.6^{+0.6+1.4}_{-0.5-1.4}$	26.9	4.4 σ
W	19: 10: 40 $^{+17}_{-17}$ s 05°03'40" $^{+03/32}_{-04/55}$ "	$2.4^{+0.6+1.3}_{-0.5-1.3}$	23.4	4.0 σ

dN/dE is the γ -ray flux, RA, right ascension; Dec., declination. TS, test statistic.

Extended Data Table 2 | Dependence of measured HAWC flux at 20 TeV on spectral assumption, assuming a power law in energy parameterized by a spectral index and an exponential cutoff parameterized by E_{cutoff}

E_{cutoff}	dN/dE at 20 TeV [$\times 10^{-16}$ TeV $^{-1}$ cm $^{-2}$ s $^{-1}$]					
	Index: -1.5		Index: -2.0		Index: -2.5	
	East Lobe	West Lobe	East Lobe	West Lobe	East Lobe	West Lobe
No cutoff	$1.0^{+0.3}_{-0.2}$	$0.9^{+0.3}_{-0.2}$	$2.4^{+0.6}_{-0.5}$	$2.1^{+0.6}_{-0.5}$	$3.3^{+0.9}_{-0.7}$	$2.4^{+0.9}_{-0.6}$
50 TeV	$4.7^{+1.1}_{-0.9}$	$4.2^{+1.1}_{-0.9}$	$5.0^{+1.2}_{-1.0}$	$4.1^{+1.3}_{-0.9}$	$3.2^{+0.9}_{-0.7}$	$1.7^{+1.1}_{-0.7}$
300 TeV	$1.7^{+0.5}_{-0.4}$	$1.6^{+0.5}_{-0.4}$	$3.3^{+0.8}_{-0.7}$	$2.9^{+0.8}_{-0.7}$	$3.6^{+0.9}_{-0.7}$	$2.4^{+0.9}_{-0.7}$

Extended Data Table 3 | Systematic uncertainties on the flux of VHE γ -rays from SS 433 measured by HAWC

Systematic	East Lobe	West Lobe
Detector Systematic Effects		$\pm 50\%$
MGRO J1908+06 Modeling		$< \pm 20\%$
Galactic Diffuse Contamination	-10%	-20%
Total	$\pm 55\%$	$\pm 55\%$

Time-asymmetric loop around an exceptional point over the full optical communications band

Jae Woong Yoon^{1,2,7}, Youngsun Choi^{1,7}, Choloong Hahn³, Gunpyo Kim¹, Seok Ho Song^{1*}, Ki-Yeon Yang⁴, Jeong Yub Lee⁴, Yongsung Kim⁴, Chang Seung Lee⁴, Jai Kwang Shin⁴, Hong-Seok Lee⁴ & Pierre Berini^{3,5,6}

Topological operations around exceptional points^{1–8}—time-varying system configurations associated with non-Hermitian singularities—have been proposed as a robust approach to achieving far-reaching open-system dynamics, as demonstrated in highly dissipative microwave transmission³ and cryogenic optomechanical oscillator⁴ experiments. In stark contrast to conventional systems based on closed-system Hermitian dynamics, environmental interferences at exceptional points are dynamically engaged with their internal coupling properties to create rotational stimuli in fictitious-parameter domains, resulting in chiral systems that exhibit various anomalous physical phenomena^{9–16}. To achieve new wave properties and concomitant device architectures to control them, realizations of such systems in application-abundant technological areas, including communications and signal processing systems, are the next step. However, it is currently unclear whether non-Hermitian interaction schemes can be configured in robust technological platforms for further device engineering. Here we experimentally demonstrate a robust silicon photonic structure with photonic modes that transmit through time-asymmetric loops around an exceptional point in the optical domain. The proposed structure consists of two coupled silicon-channel waveguides and a slab-waveguide leakage-radiation sink that precisely control the required non-Hermitian Hamiltonian experienced

by the photonic modes. The fabricated devices generate time-asymmetric light transmission over an extremely broad spectral band covering the entire optical telecommunications window (wavelengths between 1.26 and 1.675 micrometres). Thus, we take a step towards broadband on-chip optical devices based on non-Hermitian topological dynamics by using a semiconductor platform with controllable optoelectronic properties, and towards several potential practical applications, such as on-chip optical isolators and non-reciprocal mode converters. Our results further suggest the technological relevance of non-Hermitian wave dynamics in various other branches of physics, such as acoustics, condensed-matter physics and quantum mechanics.

Chiral geometries and structures appear in a variety of natural phenomena, such as in elementary-particle spins, electromagnetic polarizations, enantiomer molecules, amino acids, and helical or spiral bio-organisms. A characteristic feature of such systems is their strong directional (or selective) response to rotational stimuli. Interestingly, chirality-induced directional responses have also been found in open physical systems operating around exceptional points^{1–7}. In these systems, complex-valued energy spectra form a chiral Riemann surface topology in control parameter space, in stark contrast to conventional systems, where chiral structures are revealed explicitly in space geometric patterns. This implicit spectral chirality results in a

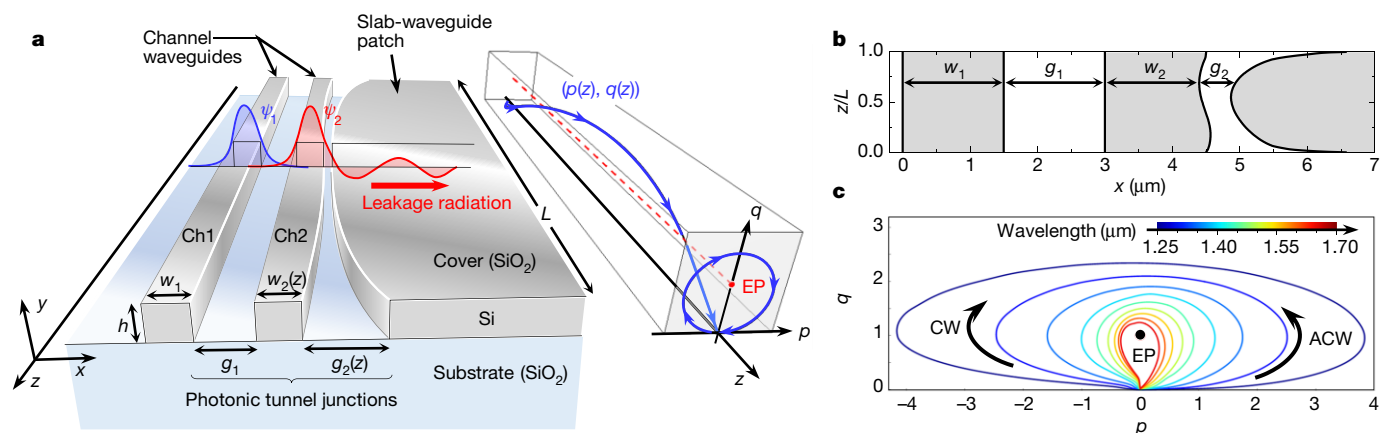


Fig. 1 | A silicon photonic architecture for dynamic EEP parametric evolution in the optical domain. **a**, Schematic of the proposed structure on a SiO₂-encapsulated Si platform. The design parameters are the Si film thickness, h , the Si core widths, w_1 and w_2 , and the photonic tunnel-barrier widths, g_1 and g_2 . For appropriately designed $w_2(z)$ and $g_2(z)$ profiles, normal modes consisting of the coupled guided photonic modes ψ_1 and ψ_2 undergo an adiabatic parametric evolution and dynamically encircle an exceptional point (EP) in the reduced-energy p - q parameter plane while propagating in the two channel waveguides (Ch1 and Ch2). The black square waveforms overlapped with the ψ_1 and ψ_2 profiles indicate

a cross-section of the Si structure in the middle of the evolution, and red arrows show the propagation direction of the leakage radiation from Ch2 into the slab-waveguide patch. The blue arrows in the right panel show the trajectories of $p(z)$ and $q(z)$ during the propagation. **b**, An optimized structure that generates EEP parametric evolution in the optical telecommunications window around 1,500 nm. Here, $h = 100$ nm. The grey and white areas indicate the waveguide-core and cladding regions, respectively. **c**, Wavelength-dependent parametric loop for the design shown in **b**. The colour bar indicates the corresponding free-space wavelength.

¹Department of Physics, Hanyang University, Seoul, South Korea. ²Electronics and Telecommunications Research Institute, Daejeon, South Korea. ³School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada. ⁴Samsung Advanced Institute of Technology, Samsung Electronics Co. Ltd., Suwon, South Korea. ⁵Department of Physics, University of Ottawa, Ottawa, Ontario, Canada. ⁶Center for Research in Photonics, University of Ottawa, Ottawa, Ontario, Canada. ⁷These authors contributed equally: Jae Woong Yoon, Youngsun Choi. *e-mail: shsong@hanyang.ac.kr

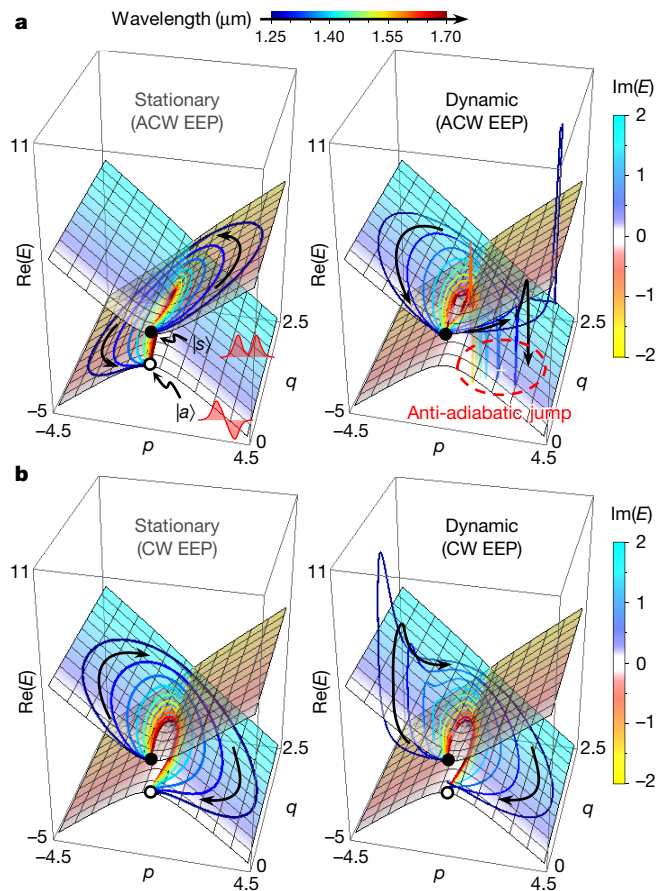


Fig. 2 | Self-intersecting eigensystem topology and time-asymmetric mode transfer. **a, b**, Stationary versus dynamic trajectory of the reduced-energy expectation value $\langle H \rangle$ for anticlockwise (ACW; **a**) and clockwise (CW; **b**) rotations along the loops indicated in Fig. 1c, on a self-intersecting Riemann sheet of the real eigenvalue $\text{Re}(E)$ of the Hamiltonian (H) in equation (1). The correspondence between wavelength and line colour is indicated in the horizontal colour bar. Here, we assume a device length of $L = 0.5$ mm. The distribution of the imaginary eigenvalue, $\text{Im}(E)$, on the $\text{Re}(E)$ surface is represented by the surface colour, as indicated in the vertical colour bar on the right. The black and white circles indicate eigenvalues of the symmetric $|s\rangle$ and antisymmetric $|a\rangle$ states, respectively, and the black arrows indicate the direction of evolution. ‘Anti-adiabatic jump’ describes an event in which the dynamic state switches eigenvalue surface.

time-asymmetric (directional) state transfer as the system dynamically steers its parametric configuration over a closed loop enclosing an exceptional point.

Directionality created in this manner may involve anomalous non-Hermitian physical effects, such as spontaneous symmetry-breaking phase transitions^{9,10}, highly deformed vector spaces¹⁷, strong imaginary energy splitting^{12–14} or breakdown of standard adiabaticity^{2–8}. Thus, further theoretical and experimental studies are required to find new wave-interaction schemes and concomitant device applications based on non-Hermitian dynamics in open systems^{18–23}.

To this end, it is of special interest to realize the encircling-an-exceptional-point (EEP) operation in the optical domain, where robust direction-selective energy or signal-transfer methods may have a substantial impact on practical applications in fields such as telecommunications, signal processing, displays, nanolithography and laser machining. However, considering the complexity of delicate time-varying non-Hermitian Hamiltonians in fabricated structures and the high precision required in configuring and controlling them^{3,4}, it is unclear whether optical EEP operations are experimentally feasible.

In this study, we demonstrate experimentally photonic modes that undergo adiabatic parametric evolutions, which dynamically encircle

an exceptional point in the reduced-energy parameter plane over the entire optical telecommunications window. This demonstration is carried out in a carefully designed silicon waveguide structure. This structure is fabricated using a single standard lithography process and is robust and completely free from practical problems arising from cryogenic operation conditions, special noise-isolation schemes and sophisticated mode-selective absorption agents, as in previous approaches^{3,4}.

The basic underlying non-Hermitian dynamics of interest here are described by a Schrödinger-type equation, $d|\psi(\tau)\rangle/d\tau = iH(\tau)|\psi(\tau)\rangle$, with a traceless binary Hamiltonian

$$H(\tau) = \begin{bmatrix} p(\tau) + iq(\tau) & 1 \\ 1 & -p(\tau) - iq(\tau) \end{bmatrix} \quad (1)$$

As reported previously^{1–5}, a Hamiltonian of this kind has a pair of parity-time-symmetric exceptional points at $p = 0$ and $q = \pm 1$ that create a self-intersecting eigenvalue topology in the parametric space represented by the time (τ)-dependent reduced-energy parameters $p(\tau)$ and $q(\tau)$. EEP parametric evolution can be obtained by suitably changing p and q over a specific time span.

To realize a robust silicon photonic structure that can be precisely controlled to obtain the desired dynamics dictated by the Hamiltonian of equation (1), we envisage a coupled-leaky-waveguide architecture, as illustrated in Fig. 1a. The structure consists of two silicon coupled-waveguide channels (Ch1 and Ch2, with corresponding core widths w_1 and w_2) and a Si slab-waveguide patch coupled with the adjacent waveguide (Ch2). In this structure, the time-dependent reduced-energy parameters $p(\tau)$ and $q(\tau)$ correspond to the real and imaginary parts of the difference between the complex propagation constants of the two channels. These are determined by the following relations:

$$\begin{aligned} p(\tau) &\approx C_p \Delta w(\tau) \\ q(\tau) &\approx C_q \exp[-\delta_{\text{clad}} g_2(\tau)] \end{aligned} \quad (2)$$

where the time parameter τ is linearly mapped onto the propagation axis z so that $\tau = \kappa z$, with κ being the inter-channel coupling constant, C_p and C_q are constant coefficients, $\Delta w(\tau) = w_2(z) - w_1$, and δ_{clad} is the decay constant of the evanescent tail of the Ch2 guided mode (see Supplementary Note I for details about the associated mathematical treatment and further explanations). The desired $p(\tau)$ and $q(\tau)$ profiles are conveniently generated by the z -dependent variation in the Ch2 core width $w_2(z)$, which adjusts the local phase velocity, and by the tunnel-barrier width profile $g_2(z)$, which independently controls the leakage radiation strength.

A trial design of Si waveguides encapsulated in SiO_2 was obtained using the full-vectorial eigenmode analysis described in Supplementary Note II. The design obtained for a fundamental transverse-magnetic (TM_{00}) mode at a Si thickness of $h = 100$ nm is illustrated in Fig. 1b and the corresponding wavelength-dependent parametric loops in the p - q plane are given in Fig. 1c. The loops clearly encircle the exceptional point at $(p, q) = (0, 1)$ over a broad spectral range from $1.25 \mu\text{m}$ to $1.70 \mu\text{m}$ that spans the entire optical telecommunications band (1.26 – $1.675 \mu\text{m}$).

Importantly, the strong persistence of the EEP loop against wavelength change is a unique property of the proposed lithographically defined structure. In Extended Data Fig. 1, we compare the wavelength-dependent parametric loop of our approach with that based on the direct index-modulation method, where the p and q profiles are directly encoded in the refractive index of the core material⁵. The latter method requires (currently unavailable) position-dependent impurity doping or deposition techniques, potentially involving dopant-solubility limits, thereby introducing chemical stability issues and atomic diffusion problems. The loop obtained with the direct index-modulation method shows strong wavelength dependence, such that the adiabaticity of the parametric change is broken in the short-wavelength region relative to the central wavelength of $1.55 \mu\text{m}$, and the loop

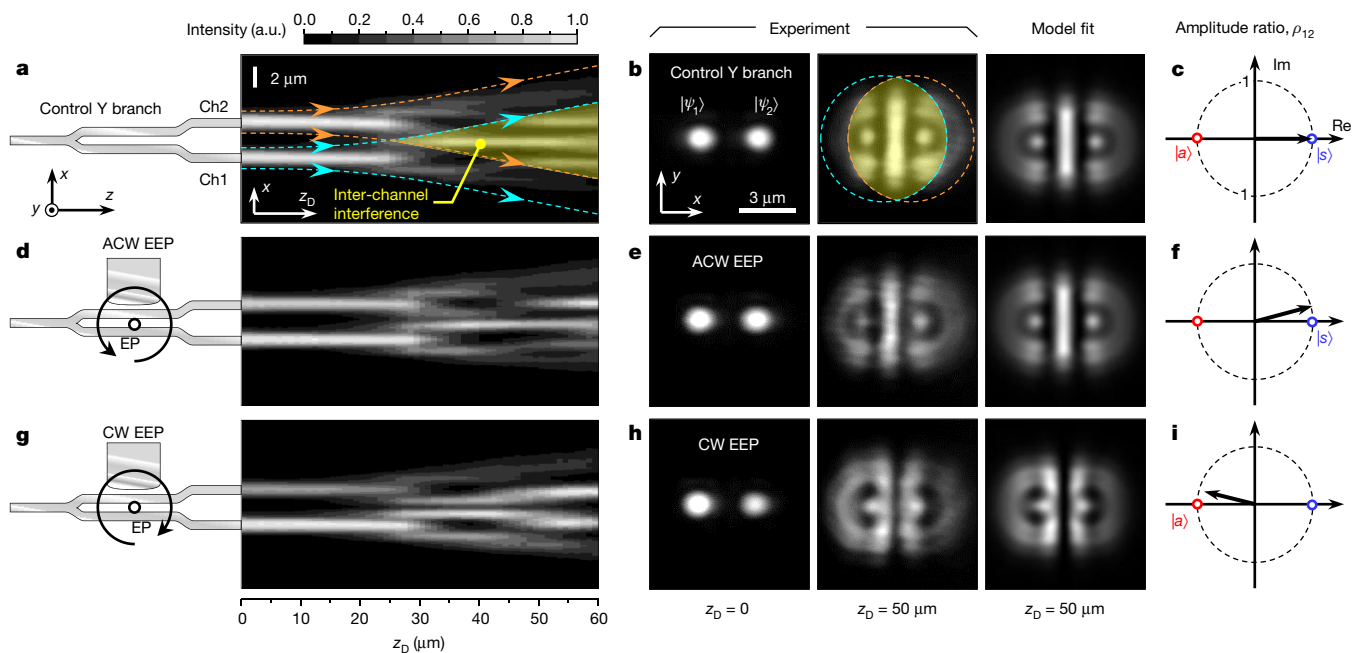


Fig. 3 | Final-state measurement based on diffraction-assisted inter-channel interference. **a–c**, Diffraction pattern analysis for a control Y-branch device. The measured Fresnel diffraction pattern in the z_D - x plane at $y = 0$ is shown in **a**, with the inter-channel interference region highlighted in yellow. The cyan and orange arrows show ray-optical representations of the diffracted output states from Ch1 and Ch2, respectively, whose beam boundaries at $z_D = 50 \mu\text{m}$ are indicated by the corresponding dashed-line circles in **b**. **b**, Diffraction patterns in

the x - y plane at $z_D = 0$ and $z_D = 50 \mu\text{m}$ are shown in the left and middle panels, respectively. The calculated diffraction pattern in the x - y plane at $z_D = 50 \mu\text{m}$, obtained using rigorous model fitting, is shown in the right panel for comparison. From the model fit, the complex amplitude ratio $\rho_{12} = a_1/a_2$ is inferred, and its phasor representation is presented in **c**. **d–f**, Diffraction pattern analysis for a ACW EEP device with $L = 0.5 \text{ mm}$. **g–i**, Diffraction pattern analysis for a CW EEP device with $L = 0.5 \text{ mm}$. The free-space wavelength of the incident light is $1.55 \mu\text{m}$ (**a–h**).

does not enclose the exceptional point in the long-wavelength region. The enhanced wavelength tolerance of the EEP loop of the proposed approach involves intricate modal-dispersive characteristics that cancel each other in the corresponding p and q parameters (see Supplementary Note I for details).

An essential consequence of dynamic EEP parametric evolution is time-asymmetric topological mode transfer as a combined effect of self-intersecting complex-eigenvalue topology and breakdown of the standard adiabaticity with strong imaginary-eigenvalue splitting^{2,5–8}. A convenient way to confirm this effect is to look into the trajectory of the expectation value $\langle H(\tau) \rangle$ on the p - q plane. In Fig. 2, we show wavelength-dependent dynamic trajectory of $\langle H(\tau) \rangle$ for the loops indicated in Fig. 1c, in comparison with the stationary trajectory that corresponds to an ideal adiabatic evolution following the instantaneous eigenvalue trajectories. These dynamic $\langle H(\tau) \rangle$ trajectories are numerically calculated using the standard Runge–Kutta method (RK4) for a model based on a modified coupled-mode formalism with matrix elements accurately determined via full-vectorial finite-element method computations. We note that the expectation value $\langle H \rangle$ here implies the c-product $\langle \psi^* | H | \psi \rangle$, following the biorthogonal treatment for non-Hermitian systems (see Supplementary Note II for details). Figure 2 clearly confirms broadband time-asymmetric topological evolution, where the dynamic passages follow symmetry-preserving anti-adiabatic trajectories for the anticlockwise (ACW) EEP, whereas the clockwise (CW) EEP results in symmetry-exchanging adiabatic passages for all wavelengths in the entire optical telecommunications band.

We fabricated the optimized devices using standard nanolithography techniques (see Methods). Optical and scanning electron microscopy were used to determine the dimensions of the devices and confirm that the desired geometric features were obtained, as shown in Extended Data Fig. 2.

In our devices, the input fundamental guided mode splits into ψ_1 and ψ_2 in a symmetrical fashion to produce a symmetric initial state $|\psi_{\text{init}}\rangle = |s\rangle = |\psi_1\rangle + |\psi_2\rangle$ that enters the EEP device. This state then undergoes the desired dynamic EEP parametric evolution, and the

final state $|\psi_{\text{final}}\rangle$ is emitted at the output facet towards free space (to the right in Fig. 3a, d, g). A key quantity for evaluating the final state $|\psi_{\text{final}}\rangle = a_1|\psi_1\rangle + a_2|\psi_2\rangle$ in the experiment is the complex amplitude ratio $\rho_{12} = a_1/a_2$, which equals 1 for the symmetric state $|s\rangle$, -1 for the antisymmetric state $|a\rangle = |\psi_1\rangle - |\psi_2\rangle$ and is a complex number otherwise.

We precisely determine ρ_{12} by analysing the diffracted intensity pattern of $|\psi_{\text{final}}\rangle$, as shown in Fig. 3. $|\psi_{\text{final}}\rangle$ emitted at the output facet is linearly diffracted in free space to form a characteristic intensity pattern containing diffraction-assisted channel-interference fringes (see Extended Data Fig. 3 for the setup used to excite the device and to measure the output diffraction pattern). The diffraction pattern shown in Fig. 3a, which was measured for a control Y-branch device (see Methods) at an excitation wavelength of $1.55 \mu\text{m}$, shows two diffracting and diverging light beams from $|\psi_1\rangle$ and $|\psi_2\rangle$ that create a characteristic interference pattern over the free-space propagation distance $z_D > 20 \mu\text{m}$, as indicated by the yellow-highlighted triangle; see Supplementary Video 1 for the continuous change of cross-sectional (x - y plane) diffracted intensity patterns as z_D progresses from 0 to $60 \mu\text{m}$. The on-focus ($z_D = 0$) and diffracted ($z_D = 50 \mu\text{m}$) cross-sectional images shown in Fig. 3b were used to determine the absolute magnitude and phase of ρ_{12} (see Supplementary Note III for details). In Fig. 3b, the experimental interference pattern at $z_D = 50 \mu\text{m}$ for the control Y-branch device shows a quantitative agreement with the theoretical interference pattern obtained using the fitted ρ_{12} value. The phasor representation of the complex amplitude ratio ρ_{12} shown in Fig. 3c corresponds exactly to the symmetric state $|s\rangle$, as expected.

This measurement method was applied to ACW and CW EEP devices with $L = 0.5 \text{ mm}$, as shown in Fig. 3d–i. From these results, we confirm that $|\psi_{\text{final}}\rangle$ approximately corresponds to the symmetric state $|s\rangle$ for the ACW EEP device and to the antisymmetric state $|a\rangle$ for the CW EEP device. Specifically in the measured cross-sectional images at $z_D = 50 \mu\text{m}$ shown in Fig. 3b, e, h, we note that the vertical antinode (bright fringe) at the centre of the patterns for the control Y-branch and ACW EEP devices is a clear signature of $|s\rangle$ (constructive

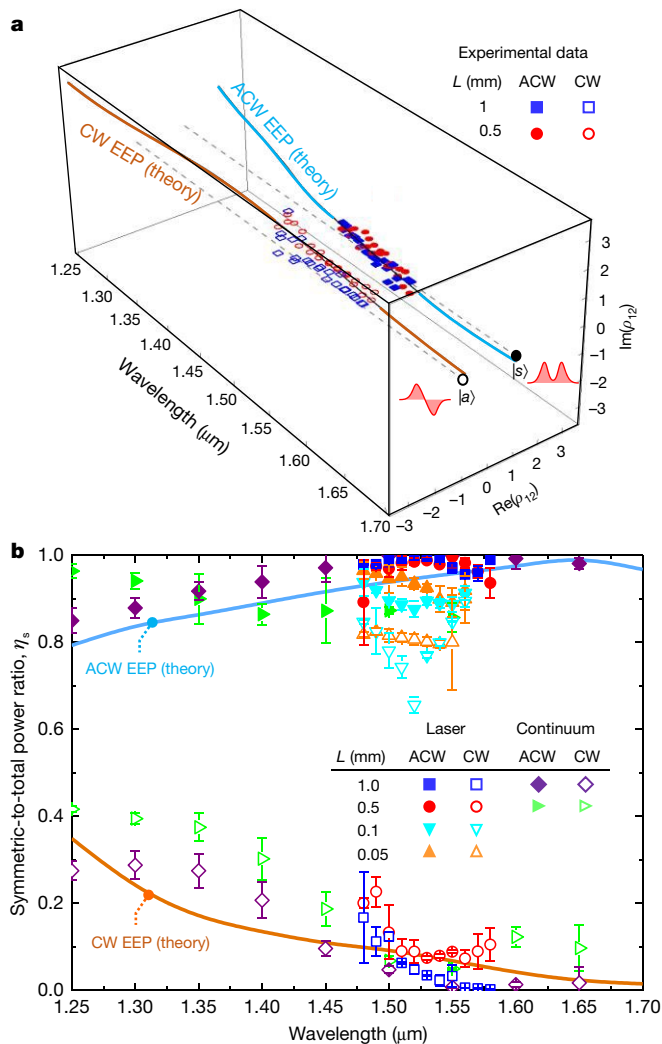


Fig. 4 | Broadband time-asymmetric state transfer. **a**, Measured complex amplitude ratio ρ_{12} as a function of free-space wavelength for ACW and CW EEP devices of length $L = 1$ mm and $L = 0.5$ mm (symbols) compared to theoretical calculations for a device length of $L = 1$ mm (solid curves). The grey dotted lines indicate the spectra for purely symmetric ($|s\rangle$) and antisymmetric ($|a\rangle$) states. **b**, Measured symmetric-to-total power ratio (symbols) for ACW and CW EEP devices of length $L = 1$ mm, 0.5 mm, 0.1 mm and 0.05 mm, compared to theoretical calculations for a device length of $L = 1$ mm (solid curves). The light sources used for the measurements were a tunable diode laser ('Laser') and a supercontinuum source ('Continuum'). The error bars indicate tolerable ranges imposed by the model fit within 10% of the least-squares error in the diffraction-assisted inter-channel interference pattern analysis.

interference), whereas the vertical node (dark fringe) at the centre of the pattern for the CW EEP device is a clear signature of $|a\rangle$ (destructive interference). Therefore, we confirm that the 0.5-mm-long EEP devices indeed produce a time-asymmetric mode-transfer effect at the excitation wavelength of $1.55\mu\text{m}$.

To further evaluate the dependence of the observed time-asymmetric mode-transfer properties on the device length, L , and the excitation wavelength, we repeated the diffraction-assisted channel interference analysis in the excitation wavelength range 1.48 – $1.58\mu\text{m}$ for several device lengths L (the excitation wavelength range is limited by the tunable laser used in the experiments). The results are summarized in Fig. 4. In Fig. 4a, we provide the measured ρ_{12} spectra for the ACW and CW EEP devices for $L = 1$ mm and $L = 0.5$ mm (see Extended Data Figs. 4, 5 for the measured and model-fitted interference patterns that give these spectral data). We confirm a clear separation between the ACW and CW cases, in quantitative agreement with the theoretical

spectra. Therefore, the time-asymmetric topological mode transfer is experimentally confirmed over a 100 -nm-wide spectral band. Notably, the consistency of the experimental spectra with the theoretical calculations further suggests that the observed effect should persist over the entire optical telecommunications band, from $1.25\mu\text{m}$ to $1.7\mu\text{m}$.

Considering the potential application of this broadband time-asymmetric effect to integrated nonreciprocal devices, the fractional power of the symmetric part in the final state is a key quantity^{3–5}. The symmetric-to-total power ratio in the final state is inferred from:

$$\eta_s = \frac{|\langle s | \psi_{\text{final}} \rangle|^2}{|\langle \psi_{\text{final}} | \psi_{\text{final}} \rangle|^2} = \frac{|1 + \rho_{12}|^2}{1 + |\rho_{12}|^2} \quad (3)$$

Figure 4b shows the experimentally obtained η_s values taken from the ρ_{12} data in Fig. 4a (see symbols labelled as 'Laser'). The values of η_s for $L = 1$ mm and $L = 0.5$ mm are persistently high (near unity) for the ACW EEP cases, whereas they remain low (near zero) for the CW EEP cases. The average ACW-to-CW extinction ratio, $[\eta_s]_{\text{ACW}}/[\eta_s]_{\text{CW}}$, over the examined spectral range of 1.48 – $1.58\mu\text{m}$ is 20.1 dB for $L = 1$ mm and 9.93 dB for $L = 0.5$ mm.

For considerably shorter devices, this value drops as the system loses adiabaticity in the z -dependent parametric change. In Fig. 4b, we also provide the η_s values for ACW and CW devices of length $L = 0.1$ mm and $L = 0.05$ mm. We find detectable separation between $[\eta_s]_{\text{ACW}}$ and $[\eta_s]_{\text{CW}}$ but their differences are substantially reduced compared to those between the longer devices, with $L = 1$ mm and $L = 0.5$ mm. This small difference for short devices is mainly due to a substantial increase of $[\eta_s]_{\text{CW}}$ from 0, which implies that the adiabatic state-flip effect for the CW EEP device is weak because the parametric change is not sufficiently slow to satisfy the adiabaticity condition. By contrast, $[\eta_s]_{\text{ACW}}$ remains high (greater than 0.8) because the fast non-adiabatic parametric change in the perturbative level does not considerably alter the initial symmetric state. Maintaining a high ACW-to-CW extinction ratio in shorter devices is an interesting challenge for on-chip device applications. Powerful heuristic optimization methods, such as genetic algorithm and particle-swarm optimization, may be applied to find better-performing parametric loops and optimal encircling-loop speed profiles for fast parametric evolution, as discussed in ref. 3.

Finally, we tested the device performance using broadband light from an infrared supercontinuum source (UWS-1000H, Santec) emitting from $1.1\mu\text{m}$ to $2.0\mu\text{m}$. In this measurement, we injected 50 -mW light from the entire continuum into the device, and the final state, emitted from the output facet into free space, was filtered using a tunable 10 -nm-wide bandpass filter at nine selected wavelengths from $1.25\mu\text{m}$ to $1.65\mu\text{m}$ in 50 -nm increments. Taking into account lensed-fibre-to-waveguide coupling losses of about 5 dB, the intensity of the excited guided mode in the device region is about 0.4 MW cm^{-2} , which demonstrates the robustness and feasibility of the proposed device as a practical broadband EEP system. The measured $[\eta_s]_{\text{ACW}}$ and $[\eta_s]_{\text{CW}}$ spectra for $L = 1$ mm and $L = 0.5$ mm plotted in Fig. 4b (see symbols labelled as 'Continuum') clearly confirm the broadband performance of the devices, which is also consistent with the theoretical model results, even though the model does not take into account minor fabrication errors and other potential factors, such as two-photon absorption and associated free-carrier effects. Measured and model-fitted interference patterns for these additional experimental data are provided in Extended Data Fig. 6.

Importantly, our measurements confirm that asymmetric modal transmission can generally persist over a very broad optical bandwidth—here, over the entire optical telecommunications window, from $1.26\mu\text{m}$ to $1.675\mu\text{m}$ (above 400 nm). Thus, further work on the realization of, for example, monolithically integrated on-chip optical isolators, is strongly motivated. Because previous approaches exploited resonant effects, a crucial limitation of those schemes is their strong wavelength selectivity and, consequently, their narrowband operation. For example, micro-cavity-based approaches yield an operating bandwidth of

less than 0.1 nm, which corresponds to the linewidth of the resonance used^{12,13,24,25}. In addition, approaches based on photonic inter-band transitions or dynamic index-modulation schemes are restricted to the bandwidth over which group-velocity matching is maintained, which is of the order of 10 nm^{26,27}. Importantly, the asymmetric light-transmission effect based on EEP operation demonstrated here persists at least over an extremely broad 400-nm-wide bandwidth and does not use any resonance-enhancement scheme or high-speed time-varying index control.

Breaking Lorentz reciprocity could be achieved in our system by introducing a gain-saturation effect or involving another optical nonlinearity. For instance, incorporating optical gain produced by rare-earth-doped materials²⁸ or exploiting nonlinearities inherent to Si²⁹—such as stimulated Raman scattering, two-photon absorption or free-carrier effects—could be achieved without modifying the device concept substantially. Moreover, our EEP strategy, which is enabled by standard nanolithography, is in principle applicable to other material systems, such as waveguides fabricated on direct-bandgap semiconductors, which are attractive materials owing to their high optical gain and nonlinearities in the infrared, as well as their ability to pump devices electrically³⁰.

In summary, we have experimentally demonstrated a robust silicon photonic device that creates broadband topological time asymmetry by dynamically encircling an exceptional point. Precise parametric controls required for the delicate non-Hermitian wave interaction are realized on the basis of intricate photonic interactions between guided modes and the two-dimensional radiation continuum on this promising silicon photonics platform. Therefore, we have established an important experimental step towards broadband non-Hermitian integrated photonics and we expect that our results will trigger further theoretical and experimental studies on novel optical phenomena and devices based on extended open-system photonic architectures.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0523-2>.

Received: 12 October 2017; Accepted: 19 July 2018;

Published online: 17 September 2018

1. Dembowski, C. et al. Experimental observation of the topological structure of exceptional points. *Phys. Rev. Lett.* **86**, 787–790 (2001).
2. Gilyar, I., Mailybaev, A. A. & Moiseyev, N. Time-asymmetric quantum-state-exchange mechanism. *Phys. Rev. B* **88**, 020102 (2013).
3. Doppler, J. et al. Dynamically encircling an exceptional point for asymmetric mode switching. *Nature* **537**, 76–79 (2016).
4. Xu, H., Mason, D., Jiang, L. & Harris, J. G. E. Topological energy transfer in an optomechanical system with exceptional points. *Nature* **537**, 80–83 (2016).
5. Choi, Y. et al. Extremely broadband, on-chip optical nonreciprocity enabled by mimicking nonlinear anti-adiabatic quantum jumps near exceptional points. *Nat. Commun.* **8**, 14154 (2017).
6. Uzdin, R., Mailybaev, A. & Moiseyev, N. On the observability and asymmetry of adiabatic state flips generated by exceptional points. *J. Phys. A* **44**, 435302 (2011).
7. Hassan, A. U., Zhen, B., Soljačić, M., Khajavikhan, M. & Christodoulides, D. N. Dynamically encircling exceptional points: exact and polarization state conversion. *Phys. Rev. Lett.* **118**, 093002 (2017).
8. Milburn, T. J. et al. General description of quasiadiabatic dynamical phenomena near exceptional points. *Phys. Rev. A* **92**, 052124 (2015).
9. Klaiman, S., Günther, U. & Moiseyev, N. Visualization of branch points in PT-symmetric waveguides. *Phys. Rev. Lett.* **101**, 080402 (2008).
10. Rüter, C. E. et al. Observation of parity–time symmetry in optics. *Nat. Phys.* **6**, 192–195 (2010).

11. Regensburger, A. et al. Parity–time synthetic photonic lattices. *Nature* **488**, 167–171 (2012).
12. Chang, L. et al. Parity–time symmetry and variable optical isolation in active-passive-coupled microresonators. *Nat. Photon.* **8**, 524–529 (2014).
13. Peng, B. et al. Parity–time symmetric whispering-gallery microcavities. *Nat. Phys.* **10**, 394–398 (2014).
14. Feng, L., Wong, Z. J., Ma, R.-M., Wang, Y. & Zhang, X. Single-mode laser by parity–time symmetry breaking. *Science* **346**, 972–975 (2014).
15. Gao, T. et al. Observation of non-Hermitian degeneracies in a chaotic exciton-polariton billiard. *Nature* **526**, 554–558 (2015).
16. Hahn, C. et al. Observation of exceptional points in reconfigurable non-Hermitian vector-field holographic lattices. *Nat. Commun.* **7**, 12201 (2016).
17. Bender, C. M., Brody, D. C., Jones, H. F. & Meister, B. K. Faster than Hermitian quantum mechanics. *Phys. Rev. Lett.* **98**, 040403 (2007).
18. El-Ganainy, R. et al. Non-Hermitian physics and PT symmetry. *Nat. Phys.* **14**, 11–19 (2018).
19. Horsley, S. A. R., Artoni, M. & La Rocca, G. C. Spatial Kramers–Kronig relations and the reflection of waves. *Nat. Photon.* **9**, 436–439 (2015).
20. Chen, W., Özdemir, Ş. K., Zhao, G., Wiersig, J. & Yang, L. Exceptional points enhance sensing in an optical microcavity. *Nature* **548**, 192–196 (2017).
21. Hodaie, H. et al. Enhanced sensitivity at higher-order exceptional points. *Nature* **548**, 187–191 (2017).
22. Assaworrorarit, S., Yu, X. & Fan, S. Robust wireless power transfer using a nonlinear parity–time-symmetric circuit. *Nature* **546**, 387–390 (2017).
23. Goldzak, T., Mailybaev, A. A. & Moiseyev, N. Light stops at exceptional points. *Phys. Rev. Lett.* **120**, 013901 (2018).
24. Bi, L. et al. On-chip optical isolation in monolithically integrated non-reciprocal optical resonators. *Nat. Photon.* **5**, 758–762 (2011).
25. Fan, L. et al. An all-silicon passive optical diode. *Science* **335**, 447–450 (2012).
26. Yu, Z. & Fan, S. Complete optical isolation created by indirect interband photonic transitions. *Nat. Photon.* **3**, 91–94 (2009); corrigendum **3**, 303 (2009).
27. Lira, H., Yu, Z., Fan, S. & Lipson, M. Electrically driven nonreciprocity induced by interband photonic transition on a silicon chip. *Phys. Rev. Lett.* **109**, 033901 (2012).
28. Kenyon, A. J. Recent developments in rare-earth doped materials for optoelectronics. *Prog. Quantum Electron.* **26**, 225–284 (2002).
29. Lin, Q., Painter, O. J. & Agrawal, G. P. Nonlinear optical phenomena in silicon waveguides: modeling and applications. *Opt. Express* **15**, 16604–16644 (2007).
30. Bhattacharya, P. *Semiconductor Optoelectronic Devices* 2nd edn (Prentice Hall, New Jersey, 1997).

Acknowledgements This research was supported in part by the Global Frontier Program of the National Research Foundation (NRF) of Korea, which is funded by the Ministry of Science, ICT & Future Planning (NRF-2014M3A6B3063708), the Basic Science Research Program (NRF-2015R1A2A2A01007553), the Presidential Post-Doc Fellowship Program (NRF-2017R1A6A3A04011896) and the Natural Sciences and Engineering Research Council of Canada (grant RGPIN-2016-04197). We thank J. Hong and Y. Ryu for their support in the spectroscopic phase-sensitive measurement.

Reviewer information Nature thanks Q. Gu, W. Heiss and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.C., J.W.Y., C.H. and S.H.S. conceived the original concept and initiated the work. J.W.Y. and Y.C. developed the theory and model. Y.C., C.H. and J.W.Y. designed the optimized device samples. K.-Y.Y. and J.Y.L. fabricated and characterized the samples with the support by Y.K., C.S.L., J.K.S. and H.-S.L. J.W.Y., G.K., C.H. and Y.C. performed the spectral measurement. J.W.Y., Y.C., G.K. and S.H.S. analysed the measurement results. All authors discussed the results. J.W.Y., Y.C., P.B. and S.H.S. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0523-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0523-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.H.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The fabrication processes used were: plasma-enhanced chemical vapour deposition (PECVD) of a 100-nm-thick amorphous Si film on a quartz substrate; electron-beam lithography to define the desired pattern on a negative electron-beam resist; anisotropic reactive ion etching to transfer the resist pattern to the Si film; resist removal; SiO₂ spin-on-glass coating to form the top cladding layer; dicing; and focused ion-beam (FIB) milling to produce clean high-quality input and output facets. Details about the processing conditions are provided below.

Si film deposition. The SiH₄-based PECVD process was carried out with a CENTURA series PECVD system (Applied Materials). The processing conditions were: chamber temperature of 350 °C, base pressure of 2 torr, power of 60 W, and SiH₄ and H₂ gas flow rates of 55 and 110 standard cubic centimetres per minute (sccm), respectively.

Electron-beam lithography. The waveguide and side-patch patterns were generated using a JBX-9500FS electron-beam lithography system (JEOL Co.). A negative electron-beam resist (OEBC-CAN038 from TOK) was spun on a Si/quartz substrate. Electron-beam writing on the resist was optimized within an exposure-dose range

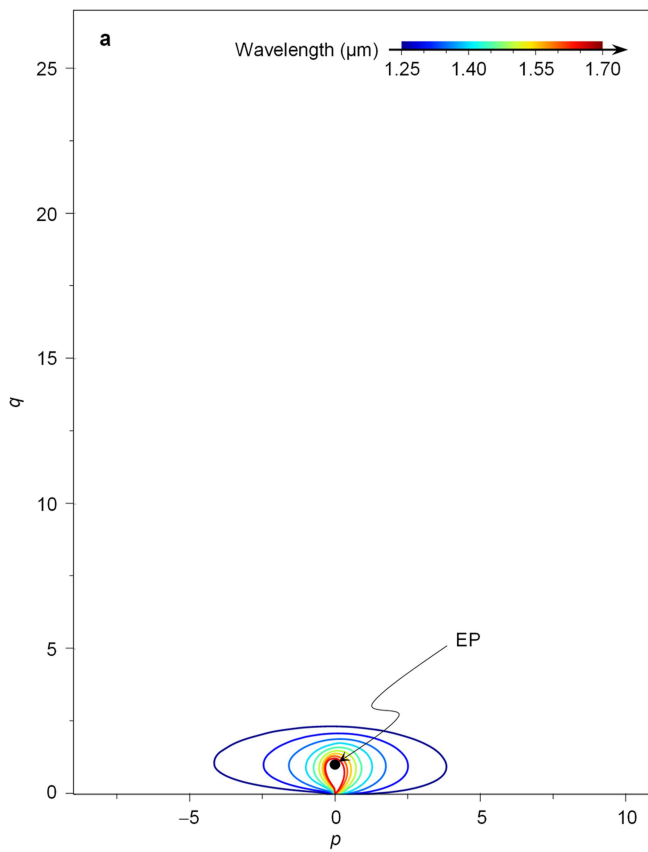
of 80–120 μC cm⁻² and a 2.38% tetramethyl ammonium hydroxide solution was used for the resist development.

Reactive ion etching. The resist pattern was transferred to the Si film using an inductively coupled plasma (ICP) reactive ion etching system from STS Co. The process was performed with a C₄F₈/SF₆/Ar gas mixture at flow rates of 45, 39 and 10 sccm. Other processing conditions were: ICP system power of 2 kW, bias power of 30 W and chamber base pressure of 10 mtorr.

SiO₂ encapsulation. As a final step, the generated Si-waveguide device patterns were encapsulated in a 1.2-μm-thick SiO₂ film using the spin-on-glass technique. Spin-on glass (AZ LEXP-S10-002 from Merk Inc.) was spin-coated on the sample, followed by post-processing baking on a hot plate at 180 °C for 1 h.

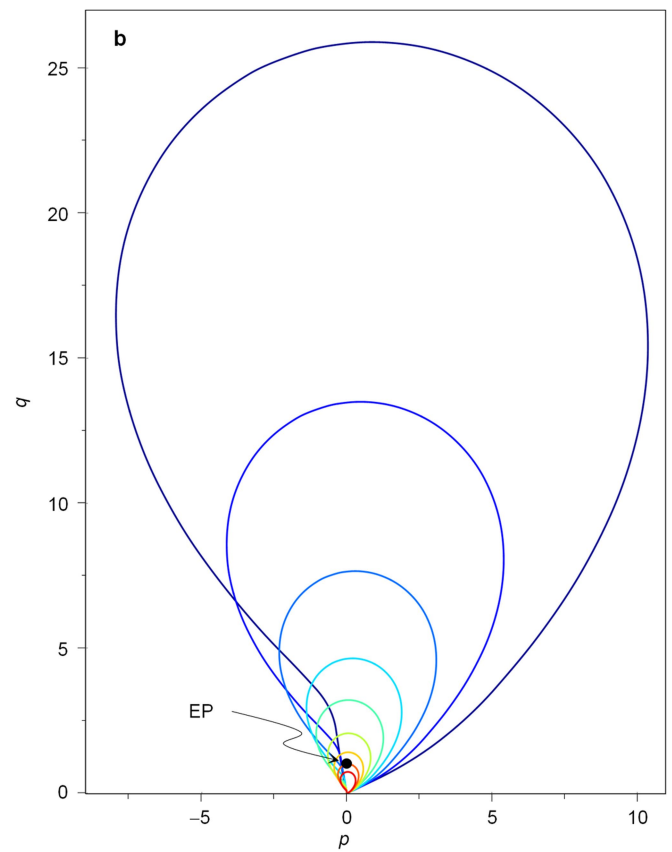
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

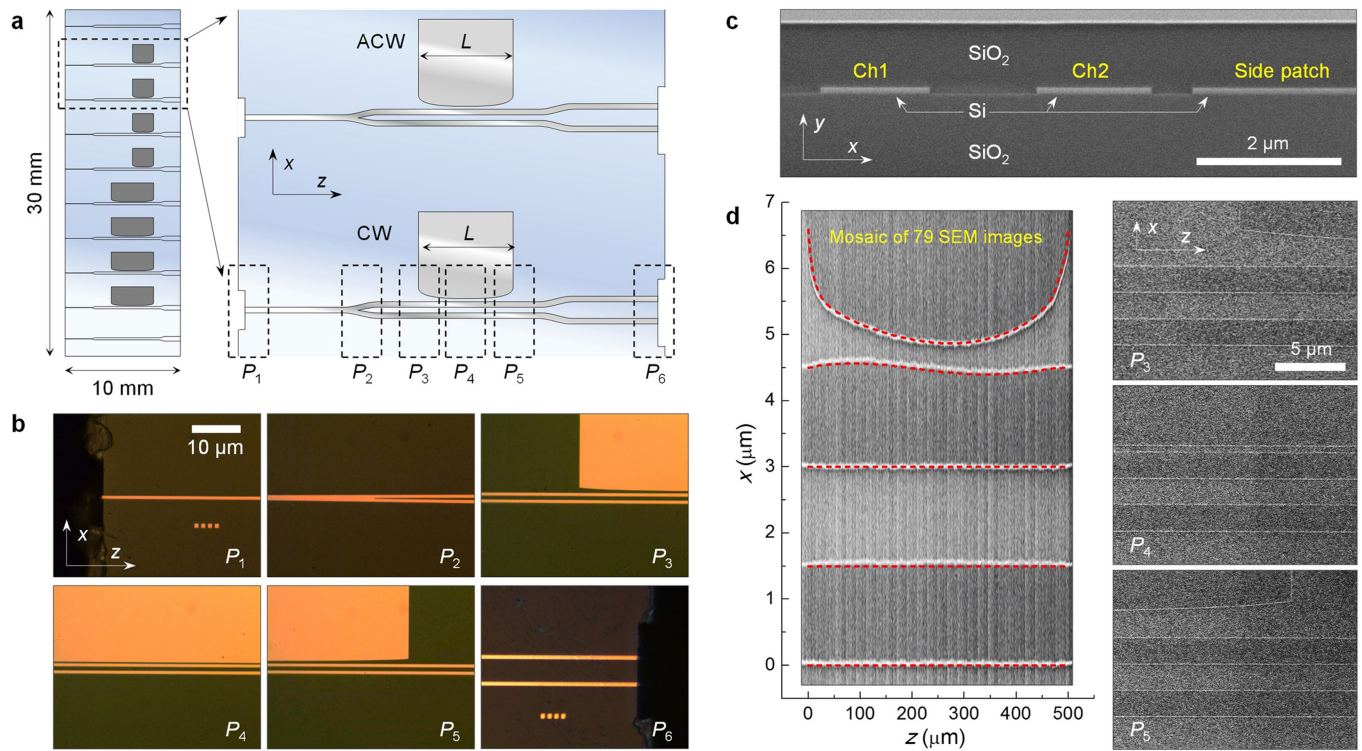


Extended Data Fig. 1 | Wavelength dispersion of an EEP loop.

a, Wavelength-dependent EEP loop obtained with the proposed approach, which is based on controlling the waveguide width and photonic tunnel gap. The line colour indicates the corresponding free-space wavelength, as shown in the colour bar. EP, exceptional point. **b**, Wavelength-dependent

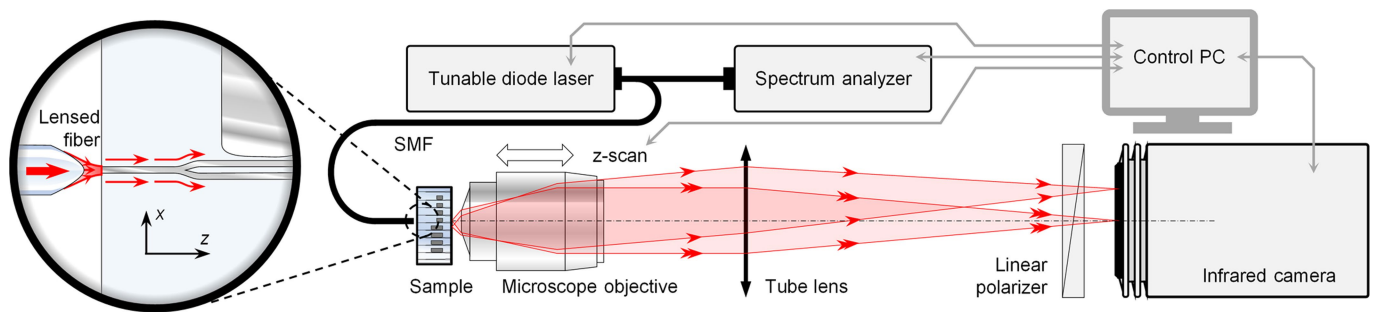


EEP loop obtained with the direct index-modulation method. For a fair comparison, the Ch1 waveguide is invariant along the z axis, the direct index-modulation profile is applied only to the Ch2 waveguide, and the profile is designed such that the loop at $1.55\ \mu\text{m}$ is identical to that in **a** at the same wavelength.



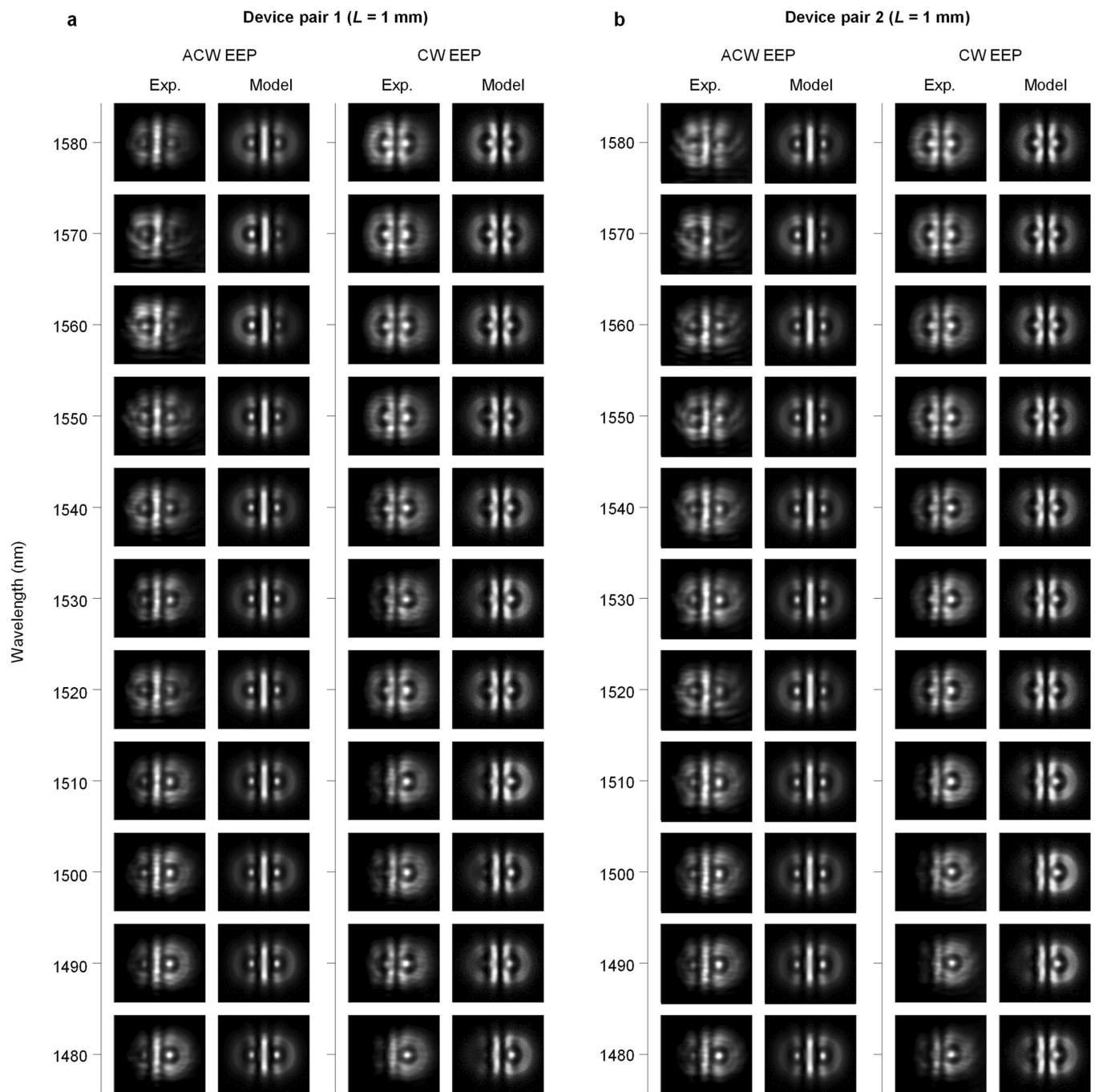
Extended Data Fig. 2 | Fabrication of SiO₂-encapsulated Si devices for dynamic EEP. **a**, Schematic of a device array fabricated on a 10 mm × 30 mm quartz substrate. As shown in the left panel, the array consists of two control Y-branch waveguides at the top and bottom of the layout and four pairs of Si-waveguide devices in the middle (pairs 1–4, from bottom to top; see Extended Data Figs. 4–6). The control Y-branch devices have two identical straight waveguides and do not include a slab-waveguide side-patch. Each Si-waveguide device pair consists of two devices of the same length L , one on top for ACW rotation and the other at the bottom for CW rotation in the p – q plane. Each device includes an FIB-prepared input

facet (in region P_1 ; see right panel), a symmetric Y-branch (in P_2), the EEP region (spanning P_3 – P_5) and an FIB-prepared output facet (in P_6). **b**, Optical microscope images of the characteristic device sections P_1 – P_6 for a ACW EEP device with $L = 0.5$ mm. **c**, **d**, Cross-sectional (**c**) and top-view (**d**) scanning electron microscope (SEM) images of a sacrificial device (ACW EEP with $L = 0.5$ mm) fabricated under the processing conditions detailed in Methods. In **d**, the left panel shows an image of the entire EEP device region, obtained by forming a mosaic of 79 successive SEM images, with the optimized design indicated by the red dashed curves. The right panels show close-up SEM images of P_3 , P_4 and P_5 .



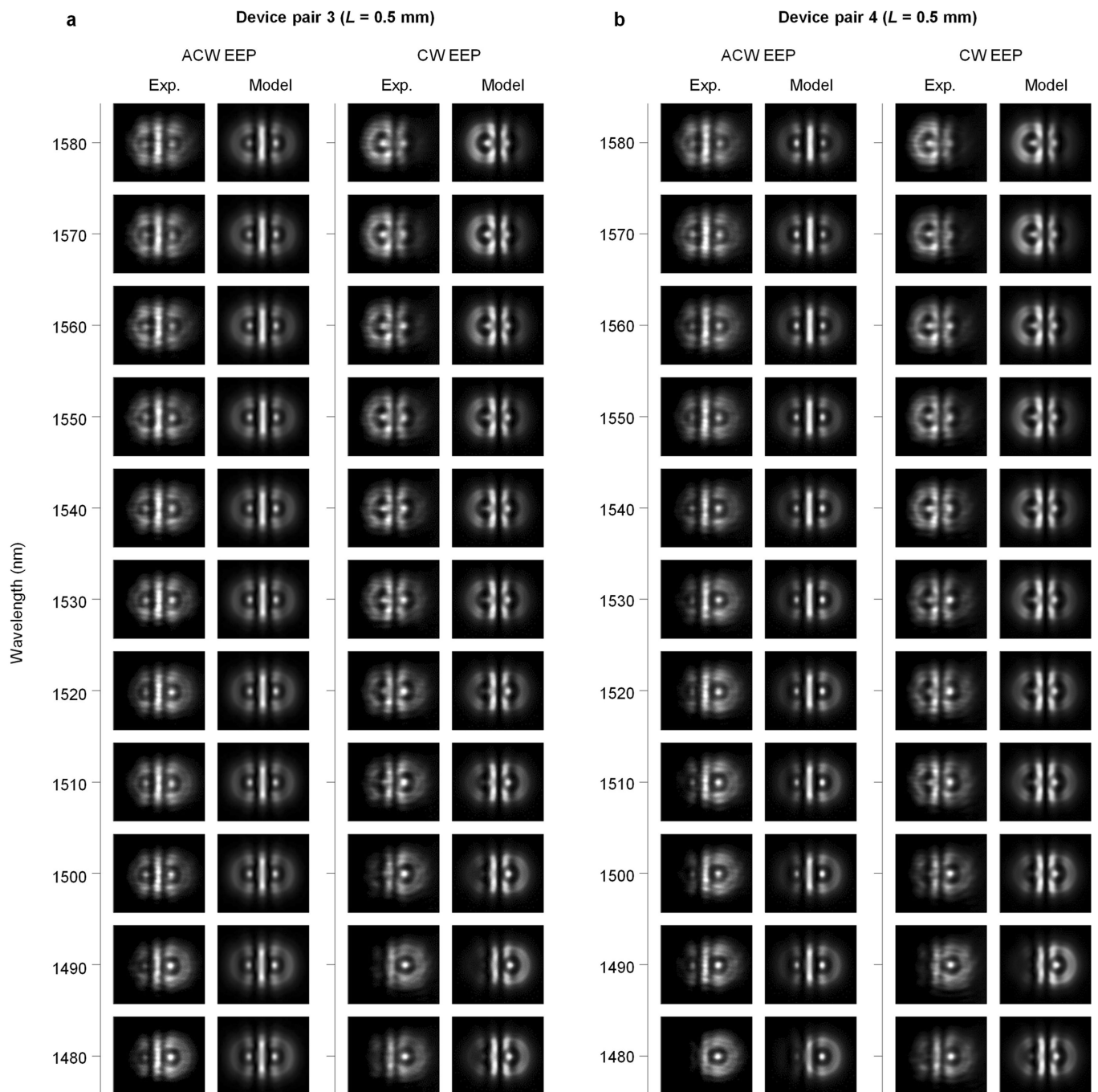
Extended Data Fig. 3 | System used for diffraction-assisted channel interference and spectral response measurements. Coherent infrared light from a tunable diode laser (MG9638A, Anritsu) is coupled to the device through a lensed single-mode fibre at the input facet. The actual excitation wavelength is monitored with a spectrum analyser (MS9710B, Anritsu). For precise alignment and stable input coupling, the lensed fibre is mounted on a three-axis translation-stage assembly and the sample is mounted on a six-axis translation-rotation-stage assembly. The output light is collected by a z-scanning microscope objective (numerical aperture, 0.3; working distance, 10 mm; from Nacet) imaging tube lens

and an infrared camera (Model 7290A Micron Viewer, Electrophysics) to acquire z -dependent diffraction-pattern images. A broadband linear polarizer (LPNIR100-MP2, Thorlabs) is located in front of the infrared camera to filter the transverse electric component (E_x) out of the acquired image. A motorized linear stage (M-505.4PD, Physik Instrumente) with 0.25- μm unit-step resolution is used for position control of the z -scanning microscope objective. The tunable diode laser, spectrum analyser, motorized linear stage and infrared camera are connected to a desktop computer that controls them in a programmed order and integrates multiple camera images for high-quality data acquisition on demand.



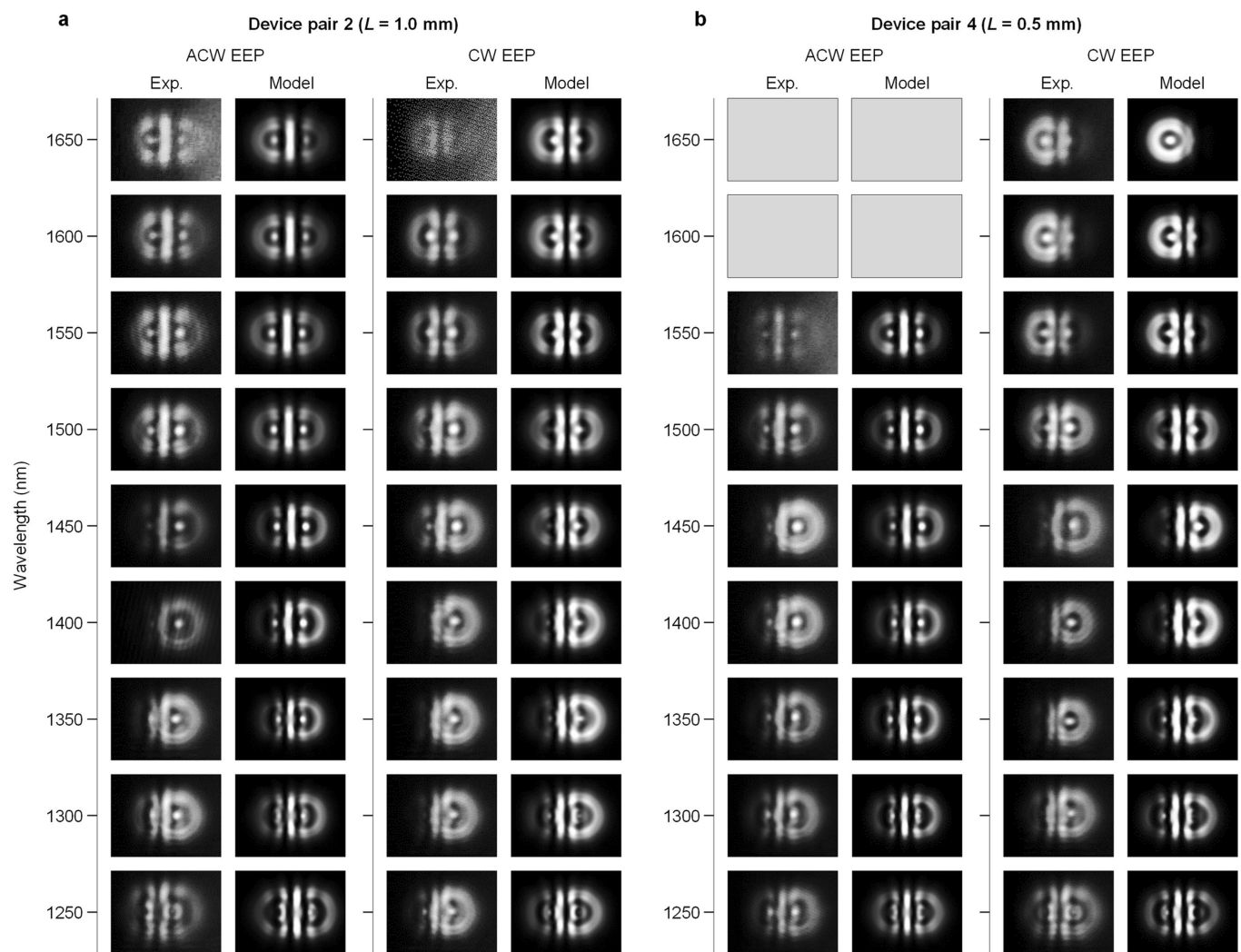
Extended Data Fig. 4 | Measured inter-channel interference patterns for device pairs 1 and 2 compared with model-fitted patterns. a, Patterns for ACW–CW EEP device pair 1 (see Extended Data Fig. 2a). **b,** Patterns for ACW–CW EEP device pair 2. The device length is $L = 1$ mm for both

device pairs 1 and 2. In the acquisition of the experimental diffraction patterns, we integrate 100 camera images to remove parasitic low-frequency electrical noise corresponding to 60-Hz harmonics. A tunable diode laser was used as the excitation source in this measurement.



Extended Data Fig. 5 | Measured inter-channel interference patterns for device pairs 3 and 4 compared with model-fitted patterns. a, Patterns for ACW–CW EEP device pair 3 (see Extended Data Fig. 2a). **b,** Patterns for ACW–CW EEP device pair 4. The device length is $L = 0.5$ mm for both

device pairs 3 and 4. In the acquisition of the experimental diffraction patterns, we integrate 100 camera images to remove parasitic low-frequency electrical noise corresponding to 60-Hz harmonics. A tunable diode laser was used as the excitation source in this measurement.



Extended Data Fig. 6 | Measured inter-channel interference patterns compared with the model-fitted patterns for device pairs excited by an infrared supercontinuum source. a, Patterns for ACW–CW EEP device pair 2 ($L = 1.0$ mm; see Extended Data Fig. 2a). **b,** Patterns for ACW–CW EEP device pair 4 ($L = 0.5$ mm). In the acquisition of the experimental

diffraction patterns, we integrate 100 camera images to remove parasitic low-frequency electrical noise corresponding to 60-Hz harmonics. In **b**, the experimental patterns at 1,600 nm and 1,650 nm for the ACW EEP device are missing owing to the very low signal-to-noise ratio at those conditions.

Giant and anisotropic many-body spin-orbit tunability in a strongly correlated kagome magnet

Jia-Xin Yin^{1,11}, Songtian S. Zhang^{1,11}, Hang Li^{2,11}, Kun Jiang³, Guoqing Chang¹, Bingjing Zhang⁴, Biao Lian⁵, Cheng Xiang^{6,7}, Ilya Belopolski¹, Hao Zheng¹, Tyler A. Cochran¹, Su-Yang Xu¹, Guang Bian¹, Kai Liu⁴, Tay-Rong Chang⁸, Hsin Lin⁹, Zhong-Yi Lu⁴, Ziqiang Wang³, Shuang Jia^{6,7}, Wenhong Wang² & M. Zahid Hasan^{1,10*}

Owing to the unusual geometry of kagome lattices—lattices made of corner-sharing triangles—their electrons are useful for studying the physics of frustrated, correlated and topological quantum electronic states^{1–9}. In the presence of strong spin-orbit coupling, the magnetic and electronic structures of kagome lattices are further entangled, which can lead to hitherto unknown spin-orbit phenomena. Here we use a combination of vector-magnetic-field capability and scanning tunnelling microscopy to elucidate the spin-orbit nature of the kagome ferromagnet Fe₃Sn₂ and explore the associated exotic correlated phenomena. We discover that a many-body electronic state from the kagome lattice couples strongly to the vector field with three-dimensional anisotropy, exhibiting a magnetization-driven giant nematic (two-fold-symmetric) energy shift. Probing the fermionic quasi-particle interference reveals consistent spontaneous nematicity—a clear indication of electron correlation—and vector magnetization is capable of altering this state, thus controlling the many-body electronic symmetry. These spin-driven giant electronic responses go well beyond Zeeman physics and point to the realization of an underlying correlated magnetic topological phase. The tunability of this kagome magnet reveals a strong interplay between an externally applied field, electronic excitations and nematicity, providing new ways of controlling spin-orbit properties and exploring emergent phenomena in topological or quantum materials^{10–12}.

Understanding and manipulating correlated quantum materials are prerequisites for exploring their potential for applications^{10–12}, and quantum materials that exhibit a giant response in the presence of an external field are particularly promising^{10–12}. Kagome antiferromagnets are central in the search for exotic quantum states because both the spin and the charge are frustrated geometrically, enabling the formation of spin-liquid phases and topological electronic structures^{1–9}. However, the realization of such states in real materials has been limited. Kagome ferromagnets are also of great interest because their unusual physics can be probed in transport, such as in transition-metal stannides^{13–18}. Transport measurements in this family have demonstrated large anomalous Hall effects that can arise from non-trivial electronic topology with non-vanishing Berry curvatures^{13–18}. The Berry phase of the antiferromagnet Mn₃Sn is associated with a non-collinear spin texture and the existence of topological fermions in its band structure^{15–17}. For the soft ferromagnet Fe₃Sn₂, it is speculated that the Berry phase is associated with a massive Dirac band near the corner of the Brillouin zone, which hosts a two-dimensional gap¹⁸. Accordingly, this family serves as a fertile platform for exploring the interplay between magnetism and quantum electronic structure in kagome lattices. Here we study the atomically resolved electronic structure of Fe₃Sn₂ at 4.2 K (Curie temperature, $T_{\text{Curie}} = 670$ K) by

using a combination of time-reversal-breaking vector-magnetic-field capability and high-resolution scanning tunnelling microscopy (STM) and spectroscopy (STS). Although many previous works focused on the unusual transport properties of Fe₃Sn₂, we observe an unexpected giant anisotropic vector-field response of the electronic states of the kagome lattice, which opens up the opportunity to demonstrate controlled quantum-level manipulation of an exotic topological phase. The methodology described here offers a new way of discovering magnetic topological phases in a strongly correlated setting, which can be used for the discovery of other correlated topological materials.

Fe₃Sn₂ has a layered rhombohedral crystal structure with space group $R\bar{3}m$ and hexagonal lattice constants $a = 5.3$ Å and $c = 19.8$ Å. It consists of a honeycomb Sn layer sandwiched between kagome FeSn bilayers (Fig. 1a). Owing to the weak bonding of these bilayers, the sample tends to cleave with either a FeSn- or a Sn-terminated surface. These two surfaces are identified experimentally via comparisons of their respective step edge heights in the crystal structure (Fig. 1b). Mapping the differential conductance of these two surfaces also reveals differences in the electronic structure (Fig. 1c). A detailed inspection of Fig. 1d confirms the honeycomb lattice structure of the Sn surface, whereas the FeSn surface exhibits smaller corrugation, hindering direct atomic identification. By analysing the line-cuts across the step edge of the Sn surface (Fig. 1e), we determine the position of the Sn atom that corresponds to the centre of the Sn honeycomb unit and those of the Fe atoms on the FeSn surface (Fig. 1d).

Having identified the two surfaces, we study the quasiparticle excitations under the perturbation of an external magnetic field. At zero field, the low-energy differential conductance spectra of both surfaces show high-intensity states around the Fermi energy, with the spectrum of the FeSn surface exhibiting an additional state at -8 mV. Increasing the c -axis field causes a pronounced shift of the side peak towards negative energies, whereas there is no discernible shift of the states near the Fermi energy. From the observed magnetic response and its surface dependence, it is likely that the side peak arises from the magnetic Fe orbital in the kagome lattice. We find that the magnetic field response of the lattice extends beyond the Zeeman effect in several aspects. The shift of the side peak saturates around 1 T, with a total energy shift of 12 meV (Fig. 2a). We observe an identical shift when the 1 T field is reversed. More importantly, the saturation behaviour agrees well with the magnetization curve (Fig. 2c), denoting a magnetization-driven shift. If this energy shift was attributed to the Zeeman effect, it would amount to an anomalously large g factor, which has not been reported in the literature (Fig. 2b).

To explore the magnetization response of Fe₃Sn₂ in three dimensions, we rotate the external field in the a - b plane. We find that this state of the FeSn surface also saturates before 1 T when the field is applied

¹Laboratory for Topological Quantum Matter and Advanced Spectroscopy (B7), Department of Physics, Princeton University, Princeton, NJ, USA. ²Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing, China. ³Department of Physics, Boston College, Chestnut Hill, MA, USA. ⁴Department of Physics and Beijing Key Laboratory of Opto-electronic Functional Materials and Micro-nano Devices, Renmin University of China, Beijing, China. ⁵Princeton Center for Theoretical Science, Princeton University, Princeton, NJ, USA. ⁶International Center for Quantum Materials and School of Physics, Peking University, Beijing, China. ⁷CAS Center for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing, China. ⁸Department of Physics, National Cheng Kung University, Tainan, Taiwan. ⁹Institute of Physics, Academia Sinica, Taipei, Taiwan. ¹⁰Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ¹¹These authors contributed equally: Jia-Xin Yin, Songtian S. Zhang, Hang Li. *e-mail: mzhhasan@princeton.edu

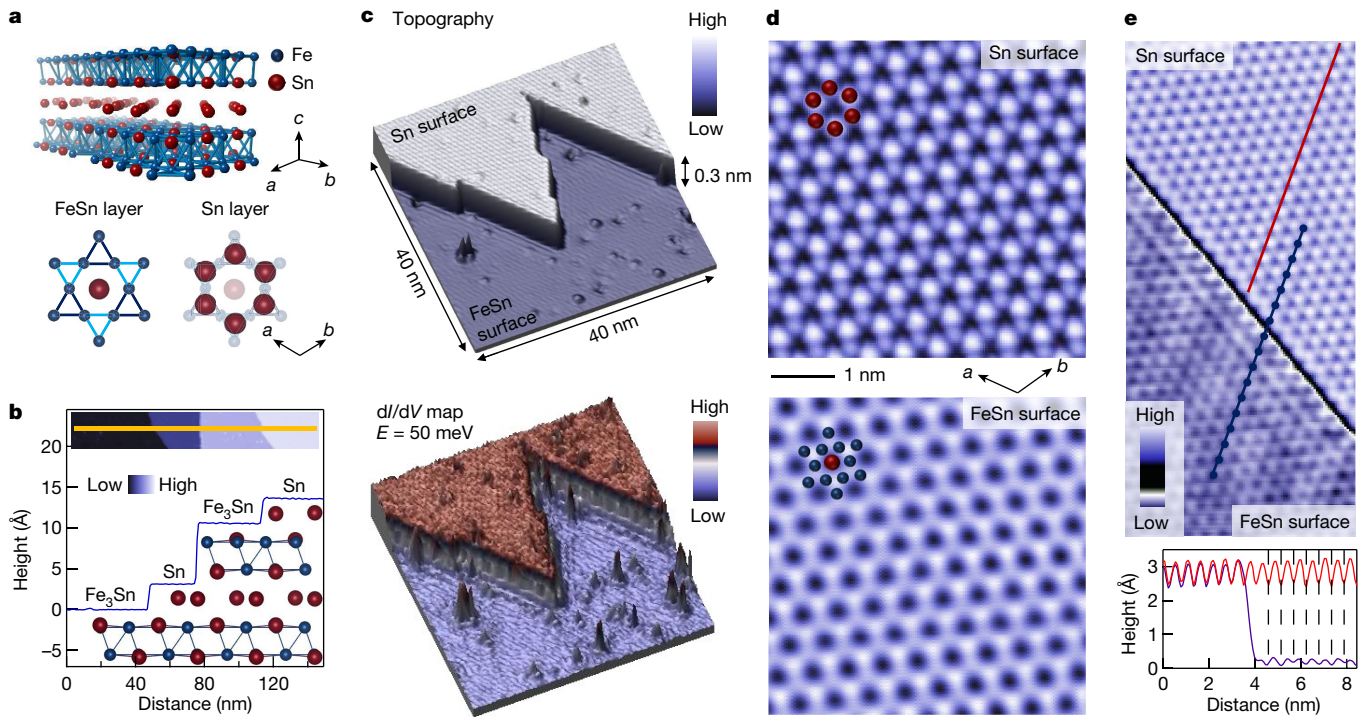


Fig. 1 | Surface identification at the atomic scale. **a**, Crystal structure of Fe_3Sn_2 . The lower illustrations show the kagome lattice of the FeSn layer (left) and the honeycomb lattice of the Sn layer (right). **b**, Atomic steps created by cryogenic cleaving. By comparing the step-edge height with the c -axis crystal structure we can determine the FeSn and Sn surfaces. The inset shows a topographic image with multiple steps and the colour bar denotes the height of the surface. The main figure shows a line-cut profile of the inset. Tunnelling junction setup: $V = 50$ mV, $I = 0.03$ nA. **c**, Topographic image of a single atomic step (top) and the corresponding

differential conductance map obtained at a bias voltage of 50 mV (bottom). **d**, Atomically resolved Sn and FeSn surfaces. The insets show the corresponding atomic lattice structures ($V = 50$ mV, $I = 0.8$ nA). **e**, Lattice alignment between the Sn surface and the FeSn surface from an edge of a Sn surface step. The lower panel compares the line-cut profiles taken along the red and blue lines in the upper image. The blue dots in the upper image and the black dashed lines in the lower plot indicate the lattice unit distance.

in plane and that the saturated shift at 1 T evolves with the azimuth angle, θ (Fig. 2d). In contrast to the six-fold crystal symmetry of Fe_3Sn_2 , this evolution has a two-fold symmetry, which can be described by the

function $3.2 - 3.2\cos(2\theta)$ meV (Fig. 2e). Notably, there is no shift when the field is applied along the a axis ($\theta = 0$). As the net magnetization is known to lie in plane at low temperatures in zero field^{19,20}, such nodal

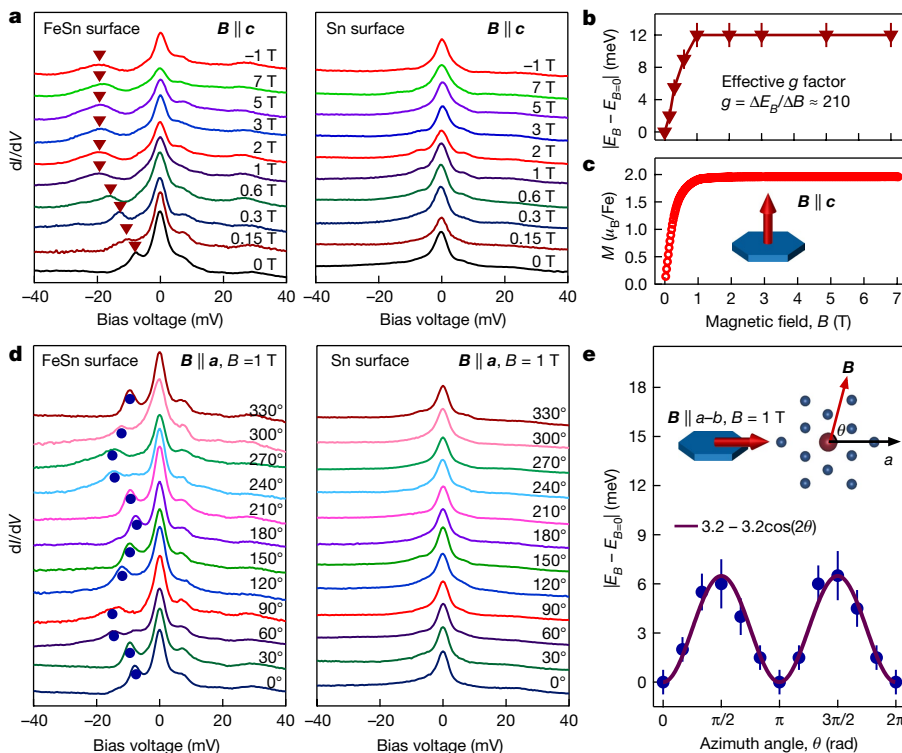


Fig. 2 | Vector-magnetization-induced giant and nematic energy shift. **a**, Dependence of the differential spectra of the FeSn and Sn surfaces on a magnetic field parallel to the c axis. The spectra are offset for clarity and the red arrows mark the peak positions. **b**, Energy shift of the electronic state of the FeSn surface as a function of the c -axis field. From the initial shift rate below 1 T, we derive an effective g factor of around 210. The error bars are based on the energy resolution. **c**, Bulk c -axis magnetization curve, expressed in units of the Bohr magneton, μ_B , per Fe atom. The magnetization correlates strongly with the energy shift. The inset shows that the field is applied perpendicular to the sample surface. **d**, Dependence of the differential spectra of the FeSn and Sn surfaces on the angle of an in-plane field ($B = 1$ T). The blue circles mark the peak positions. **e**, Energy shift as a function of azimuth angle, θ . The data can be fitted by a two-fold-symmetric function as $3.2 - 3.2\cos(2\theta)$. The inset illustrates the field's azimuth angle with respect to the a axis of the kagome lattice. The error bars are based on the energy resolution.

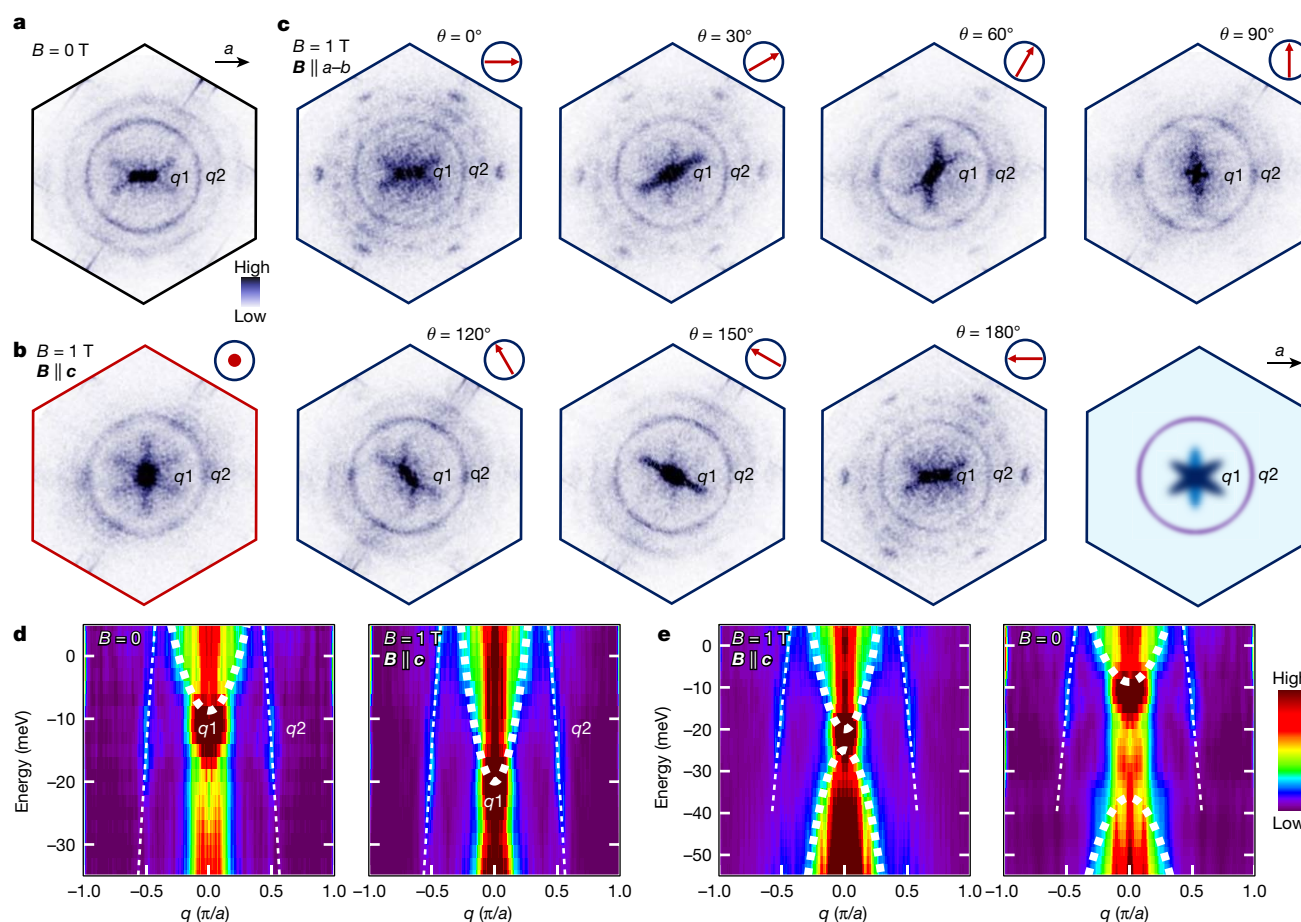


Fig. 3 | Vector-magnetization-governed electronic symmetry. **a**, QPI data of the FeSn surface obtained at the side-peak energy at zero field, showing a ring-like signal at the wave vector q_2 and a two-fold-symmetric pattern along the a axis at the wave vector q_1 . **b**, QPI data taken with a 1-T field parallel to the c axis, showing a six-fold-symmetric pattern at q_1 . **c**, QPI data taken with a 1-T in-plane vector field, showing systematic rotation of the pattern at q_1 . The schematic at the bottom right illustrates the dominant scattering vectors of the observed QPI signals. The red arrows illustrate the direction of the applied field, and the black arrow

marks the direction of spontaneous magnetization. **d**, Field-dependent QPI dispersion along the a axis (which is parallel to the Γ -M cut). The distance along the a axis is expressed in reciprocal space (π/a). **e**, QPI dispersion along the a axis with a wider energy range, showing signatures of both the upper and lower branches of the massive Dirac band. The thin dashed lines indicate the dispersion of the field-independent hole-like band (q_2). The thick dashed lines denote a possible field-dependent massive Dirac dispersion, which heuristically fits to the intensity decrease of the QPI signal around $q = 0$. All the QPI data are unsymmetrized.

behaviour can be understood by considering the spontaneous magnetization to be along the a axis, which already saturates the energy shift. The anisotropic evolution that we observe in our STM data also agrees qualitatively with the bulk transport anisotropy in response to a vector field (Extended Data Fig. 3), consistently demonstrating the existence of electronic nematicity in Fe_3Sn_2 , which has been previously observed in correlated materials^{21–24}.

To further study the symmetry of the electronic state realized in this material, we map the differential conductance of the FeSn surface over a large area under various vector-field conditions. By taking their Fourier transforms, we obtain the quasiparticle interference (QPI) data for the electron scattering involving the band structure. The QPI data obtained at the energy of this side-peak state are shown in Fig. 3. The zero-field QPI data in Fig. 3a exhibit ring-like signals at the larger wave vectors (q_2) and a two-fold pattern around the zone centre (q_1). This spontaneously broken symmetry state also aligns with the sample's a axis, in agreement with the aforementioned nematicity, which is consistent with our transport data. Remarkably, we find that the c -axis magnetization removes the electronic nematicity and restores the rotational symmetry (Fig. 3b). Although the a -axis magnetization retains the same nematic pattern at q_1 (Fig. 3c), this pattern is systematically rotated by the rotation of in-plane magnetization. When the external field is withdrawn, the nematicity recovers to that shown in Fig. 3a, regardless of magnetization history. This suggests that there exists an intrinsic

nematic order pinning the spontaneous magnetization along the a axis. By contrast, the QPI around q_2 remains approximately isotropic regardless of the field, indicating that the shifting of the electronic state is probably associated with the states near q_1 .

Such an association is further supported by the field-dependent QPI dispersion plotted in Fig. 3d, e. The QPI dispersions in Fig. 3d shows a clear hole-like band (q_2) with no discernible field dependence, corresponding to the ring-like signal seen in the QPI images in Fig. 3a–c. On the other hand, at small q , the dispersion reveals a maximum in the QPI intensity at approximately the energy of the side peak both at zero field ($E = -10 \pm 2$ mV) and at $B = 1$ T ($E = -20 \pm 2$ mV). This QPI intensity extends to larger q with increasing energy, suggesting an electron-like band with a band bottom at this low- q maximum-energy peak (q_1). Moreover, when the energy window of the dispersion is extended to even lower energies (Fig. 3e), a second hole-like band appears at $B = 1$ T, with two (one upper and one lower) branches forming a non-monotonically dispersing feature with an hourglass-like shape. The non-monotonic shape of the observed signal indicates that the scattering is sensitive to the details of the underlying dispersion in the band structure. It is consistent with a massive Dirac-like dispersion, under the assumption that the scattering is intra-band in nature and corresponds to the Dirac feature expected around this energy in the photoemission measurement¹⁸. These two dispersive branches move farther apart from the centre of the hourglass-like feature as the

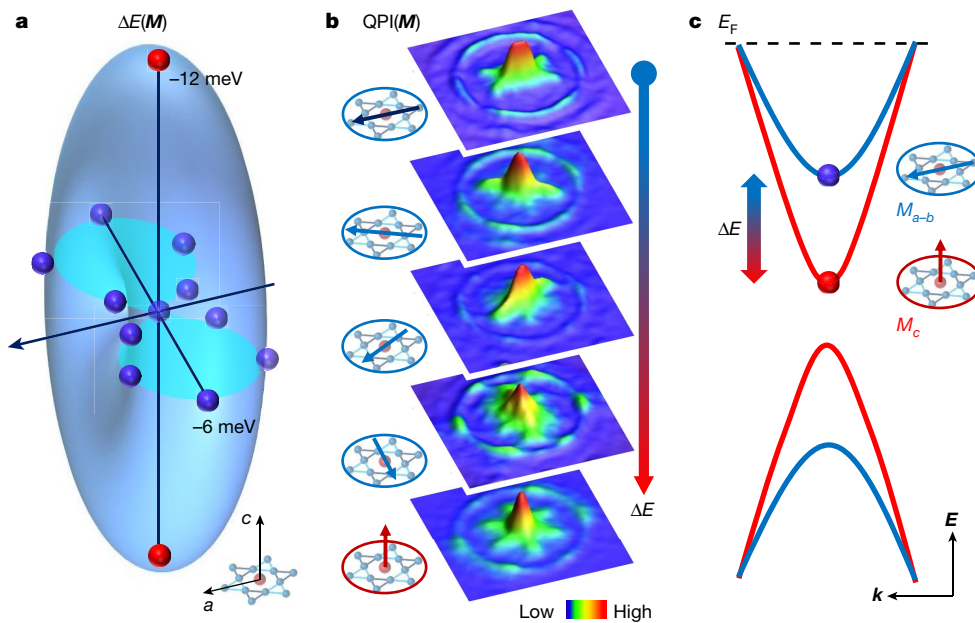


Fig. 4 | Correspondence between vector-magnetization-based energy shift and broken symmetry. **a**, Saturated energy shift $\Delta E = E_B - E_{B=0}$ of the electronic state from the kagome lattice as a function of the direction of the magnetization vector. The red and blue circles are data from Fig. 2a and d, respectively. The light-blue surface shows a three-dimensional illustration of the energy shift as a function of the magnetization vector \mathbf{M} , which exhibits a nodal line along the a axis. **b**, QPI patterns as a function of the magnetization direction, which is indicated in the insets with

respect to the kagome lattice. The uppermost QPI pattern shows the spontaneous nematicity along the a axis. Magnetization along other directions can alter, and thus control, the electronic symmetry. **c**, Schematic of the magnetization-controlled Dirac gap, with the band bottom of the upper branch corresponding to the shifting state with broken symmetry. The bands lose coherence away from the Fermi energy, E_F . The blue and red curves illustrate the massive Dirac dispersion with in-plane and out-of-plane magnetization, respectively.

magnetic field is tuned to zero. Recent photoemission-based identification of a band structure resembling massive Dirac fermions in this material suggests a gap size of about 30 mV, consistent with our data at $B = 0$ T in Fig. 3e, despite what appears to be a shift in the chemical potential, which is possibly due to surface termination effects or doping differences in the samples. It is well known that photoemission measurements lack the ability to probe the field dependence of this gap (mass of the Dirac band) or determine whether the Dirac fermions originate from the FeSn kagome lattice or the Sn honeycomb lattice, which is critical for understanding and correctly modelling the state realized in this material.

A summary of our experimental findings is shown in Fig. 4. Our results demonstrate a vector-magnetization-based energy shift of the quantum electronic states with an intriguing correspondence to symmetry breaking (Fig. 4a, b). These states form an electron band crossing the Fermi level (Fig. 4c). Without an external field, spontaneous magnetization is along the a axis; the band bottom (identified as the side peak in the tunnelling conductance in Fig. 2a) exhibits a QPI with a two-fold symmetry. Although the symmetry of the QPI rotates with the magnetization, indicating strong spin–orbit coupling (SOC) that intertwines the orbital space with the magnetic space, it is unexpected that the energy of the band bottom modulates substantially with the angle of rotation. The observed energy shift induced by in-plane magnetization also has a two-fold symmetry with its nodal line along the a axis (Fig. 4a), which is indicative of an intrinsic nematic order that pins the spontaneous magnetization direction and leads to this energy difference. Rotating the magnetization to the c axis causes the largest energy shift (Fig. 4a). These giant electronic responses driven by the magnetization direction go well beyond Zeeman physics and point to a spin–orbit-entangled, correlated magnetic topological phase, which we discuss below.

In fact, previous STM work on other systems has shown that owing to the presence of SOC, the electronic structure of magnetic thin films with domain walls²⁵ and skyrmions²⁶ can depend on the spin orientation. Because electronic structures on the kagome lattice have linear band-crossing Dirac points at the Brillouin zone corners, it is natural

for us to consider a picture of Dirac fermions in the presence of SOC in the study of kagome-lattice (quantum) anomalous Hall materials^{8,13,14}. The observed energy shift should thus result from the interplay of the Dirac gap (Fig. 4c) with magnetism. The large ferromagnetic moment splits the Dirac crossing into two branches that are well separated in energy, with spins polarized parallel and anti-parallel to the direction of magnetization. In ref. ¹⁸, the magnetization is assumed to be along the c axis, and a Kane–Mele-type²⁷ SOC that preserves the spin component S_z is considered to produce a Dirac gap. However, the spontaneous magnetization in Fe₃Sn₂ lies in plane at low temperatures, so that the S_z SOC cannot generate a gap at the Dirac crossing in the kagome lattice²⁸. This contradicts our observation of the largest mass gap (smallest energy shift) for a -axis magnetization, sketched in Fig. 4a, c. Thus, the physics governing the interplay between SOC and magnetism here lies beyond the Kane–Mele scenario. One possibility is that all SOC interactions respecting the full crystal symmetry need to be constructed with both S_z - and in-plane $S_{x,y}$ -conserving terms, and our results indicate that the latter should have a larger effect. Alternatively, the Dirac gap may have an additional source that interferes with that due to a dominant Kane–Mele SOC. Because Fe₃Sn₂ displays a large anomalous Hall effect with skyrmion excitations²⁰, it is possible that the latter has a contribution from the spin Berry phase³ associated with chiral spin textures¹⁹. The spin chirality produces a gauge flux and, according to the theory³, opens a Dirac gap independent of the magnetization direction. As the magnetization is rotated to the c axis, the orbital flux induced by the Kane–Mele SOC can be out of phase and compete with the gauge flux, leading to the reduction of the Dirac gap (Extended Data Fig. 11), consistent with our interpretation of the data.

Furthermore, our experiment reveals an intriguing nematic order in this kagome magnet. In addition to the magnetization-controlled charge nematicity due to SOC effects, there exists an intrinsic nematic order originating from the charge channel, as evidenced by the anisotropic energy shift and transport response to the vector magnetization, as well as the pinning of the spontaneous magnetization direction irrespective of vector magnetization history. Interestingly, the well known

intra-unit-cell ($q=0$) charge-ordered state on the kagome lattice, which is driven by inter-site Coulomb interactions, is a nematic state, as demonstrated theoretically^{29,30}.

In summary, our experiment uncovers a correspondence between vector-field-based energy shift and broken symmetry in Fe_3Sn_2 , which demonstrates unusually large and anisotropic magnetic tunability in a spin-orbit kagome magnet and points to an underlying correlated magnetic topological ground state. The novelty of this work is the spin-orbit tunability and the gigantic response of the kagome material, which are not implied by, or can be derived from, known transport or photoemission effects. The gigantic spin-orbit response that we discovered in this strongly correlated material is unexpected and not implied by results reported in refs^{13,14,18}. Our findings collectively show the rich and unconventional physics of kagome magnets, which encompasses entangled magnetic, charge and orbital degrees of freedom, as well as symmetry breaking and topological properties of electronic states involving low-energy fermions. A complete understanding of this physics would require a comprehensive quantum many-body theory that describes electrons in the kagome lattice in the presence of strong spin-orbit coupling. Our space-momentum exploration of electronic excitations by controlled vector-field manipulation is a powerful tool for probing the physics of topological matter beyond weakly interacting Z_2 topological insulators^{27,28}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0502-7>.

Received: 9 March 2018; Accepted: 4 July 2018;

Published online 12 September 2018.

- Mekata, M. Kagome: the story of the basketweave lattice. *Phys. Today* **56**, 12 (2003).
- Zhou, Y., Kanoda, K. & Ng, T.-K. Quantum spin liquid states. *Rev. Mod. Phys.* **89**, 025003 (2017).
- Ohgushi, K., Murakami, S. & Nagaosa, N. Spin anisotropy and quantum Hall effect in the kagomé lattice: chiral spin state based on a ferromagnet. *Phys. Rev. B* **62**, R6065 (2000).
- Yan, S., Huse, D. A. & White, S. R. Spin-liquid ground state of the $S=1/2$ kagome Heisenberg antiferromagnet. *Science* **332**, 1173–1176 (2011).
- Han, T.-H. et al. Fractionalized excitations in the spin-liquid state of a kagome-lattice antiferromagnet. *Nature* **492**, 406–410 (2012).
- Mazin, I. I. et al. Theoretical prediction of a strongly correlated Dirac metal. *Nat. Commun.* **5**, 4261 (2014).
- Chisnell, R. et al. Topological magnon bands in a kagome lattice ferromagnet. *Phys. Rev. Lett.* **115**, 147201 (2015).
- Xu, G., Lian, B. & Zhang, S.-C. Intrinsic quantum anomalous Hall effect in the kagome lattice $\text{Cs}_2\text{LiMn}_3\text{F}_{12}$. *Phys. Rev. Lett.* **115**, 186802 (2015).
- Zhu, W., Gong, S.-S., Zeng, T.-S., Fu, L. & Sheng, D. S. Interaction-driven spontaneous quantum Hall effect on a kagome lattice. *Phys. Rev. Lett.* **117**, 096402 (2016).
- Soumyanarayanan, A., Reyren, N., Fert, A. & Panagopoulos, C. Emergent phenomena induced by spin-orbit coupling at surfaces and interfaces. *Nature* **539**, 509–517 (2016).
- Keimer, B. & Moore, J. E. The physics of quantum materials. *Nat. Phys.* **13**, 1045–1055 (2017).
- Tokura, Y., Kawasaki, M. & Nagaosa, N. Emergent functions of quantum materials. *Nat. Phys.* **13**, 1056–1068 (2017).
- Kida, T. et al. The giant anomalous Hall effect in the ferromagnet Fe_3Sn_2 – a frustrated kagome metal. *J. Phys. Condens. Matter* **23**, 112205 (2011).
- Wang, Q., Sun, S., Zhang, X., Pang, F. & Lei, H. Anomalous Hall effect in a ferromagnetic Fe_3Sn_2 single crystal with a geometrically frustrated Fe bilayer kagome lattice. *Phys. Rev. B* **94**, 075135 (2016).
- Nakatsuji, S., Kiyohara, N. & Higo, T. Large anomalous Hall effect in a non-collinear antiferromagnet at room temperature. *Nature* **527**, 212–215 (2015).
- Nayak, A. K. et al. Large anomalous Hall effect driven by a nonvanishing Berry curvature in the noncollinear antiferromagnet Mn_3Ge . *Sci. Adv.* **2**, e1501870 (2016).
- Kuroda, K. et al. Evidence for magnetic Weyl fermions in a correlated metal. *Nat. Mater.* **16**, 1090–1095 (2017).
- Ye, L. et al. Massive Dirac fermions in a ferromagnetic kagome metal. *Nature* **555**, 638–642 (2018).
- Fenner, L. A., Dee, A. A. & Wills, A. S. Non-collinearity and spin frustration in the itinerant kagome ferromagnet Fe_3Sn_2 . *J. Phys. Condens. Matter* **21**, 452202 (2009).
- Hou, Z. et al. Observation of various and spontaneous magnetic skyrmionic bubbles at room temperature in a frustrated kagome magnet with uniaxial magnetic anisotropy. *Adv. Mater.* **29**, 1701144 (2017).
- Fradkin, E., Kivelson, S. A., Lawler, M. J., Eisenstein, J. P. & Mackenzie, A. P. Nematic Fermi fluids in condensed matter physics. *Annu. Rev. Condens. Matter Phys.* **1**, 153–178 (2010).
- Borzi, R. A. et al. Formation of a nematic fluid at high fields in $\text{Sr}_3\text{Ru}_2\text{O}_7$. *Science* **315**, 214–217 (2007).
- Chuang, T. M. et al. Nematic electronic structure in the “parent” state of the iron-based superconductor $\text{Ca}(\text{Fe}_{1-x}\text{Co}_x)_2\text{As}_2$. *Science* **327**, 181–184 (2010).
- Fujita, K. et al. Simultaneous transitions in cuprate momentum-space topology and electronic symmetry breaking. *Science* **344**, 612–616 (2014).
- Bode, M. et al. Magnetization-direction-dependent local electronic structure probed by scanning tunneling spectroscopy. *Phys. Rev. Lett.* **89**, 237205 (2002).
- Hanneken, C. et al. Electrical detection of magnetic skyrmions by tunnelling non-collinear magnetoresistance. *Nat. Nanotechnol.* **10**, 1039–1042 (2015).
- Kane, C. L. & Mele, E. J. Quantum spin Hall effect in graphene. *Phys. Rev. Lett.* **95**, 226801 (2005).
- Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
- Guo, H. M. & Franz, M. Topological insulator on the kagome lattice. *Phys. Rev. B* **80**, 113102 (2009).
- Nishimoto, S. et al. Metal-insulator transition of fermions on a kagome lattice at $1/3$ filling. *Phys. Rev. Lett.* **104**, 196401 (2010).

Acknowledgements Experimental and theoretical work at Princeton University was supported by the Gordon and Betty Moore Foundation (GBMF4547/Hasan) and the United States Department of Energy (US DOE) under the Basic Energy Sciences programme (grant number DOE/BES DE-FG-02-05ER46200). Work at the Institute of Physics of the Chinese Academy of Science (IOP CAS) was supported by the National Key R&D Program of China (grant number 2017YFA0206303). Work at Boston College is supported by US DOE grant DE-FG02-99ER45747. We also acknowledge the Natural Science Foundation of China (grant numbers 11790313, 11774422 and 11774424), National Key R&D Program of China (numbers 2016YFA0300403 and 2017YFA0302903), the Key Research Program of the Chinese Academy of Sciences (number XDPB08-1), Princeton Center for Theoretical Science (PCTS) and Princeton Institute for the Science and Technology of Materials (PRISM)’s Imaging and Analysis Center at Princeton University. T.-R.C. was supported by the Ministry of Science and Technology under a MOST Young Scholar Fellowship (MOST Grant for the Columbus Program number 107-2636-M-006-004-), the National Cheng Kung University, Taiwan, and the National Center for Theoretical Sciences (NCTS), Taiwan. M.Z.H. acknowledges support from Lawrence Berkeley National Laboratory and the Miller Institute of Basic Research in Science at the University of California, Berkeley in the form of a Visiting Miller Professorship. We thank D. Huse and T. Neupert for discussions.

Reviewer information Nature thanks S. Sachdev and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.-X.Y. and S.S.Z. conducted the STM and STS experiments in consultation with M.Z.H.; H.L., W.W., C.X. and S.J. synthesized and characterized the sequence of samples; K.J., G.C., B.Z., B.L., K.L., T.-R.C., H.L., Z.-Y.L. and Z.W. carried out theoretical analysis in consultation with J.Y. and M.Z.H.; I.B., T.A.C., H.Z., S.-Y.X. and G.B. contributed to sample characterization and instrument calibration; J.-X.Y., S.S.Z. and M.Z.H. performed the data analysis and figure development and wrote the paper with contributions from all authors; M.Z.H. supervised the project. All authors discussed the results, interpretation and conclusion.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0502-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.Z.H.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

STM measurements. Single crystals of Fe_3Sn_2 with sizes of up to $1\text{ mm} \times 1\text{ mm} \times 0.3\text{ mm}$ were used in this study. Samples were cleaved mechanically in situ at 77 K in ultrahigh-vacuum conditions and then immediately inserted into the STM head, which was already at He-4 base temperature (4.2 K). Before applying the magnetic field, the tip was withdrawn to about $10\text{ }\mu\text{m}$ from the sample. The vector magnetic field was applied with the zero-field cooling technique, after which we carefully approached the tip to the sample to find the same atomic-scale area to perform tunnelling spectroscopy. Tunnelling conductance spectra were obtained with an Ir–Pt tip using standard lock-in amplifier techniques with a root-mean-square oscillation voltage of 0.5 meV and a lock-in frequency of 973 Hz. The conductance maps were taken with the tunnelling junction set up $V = -50\text{ mV}$ and $I = 200\text{ pA}$, and the tunnelling spectra were taken with the junction set up $V = -100\text{ mV}$ and $I = 0.8\text{ nA}$.

Sample preparation. Single crystals of Fe_3Sn_2 were synthesized by the Sn-flux method with a molar ratio of $\text{Fe}:\text{Sn} = 1:19$. Fe and Sn grains were placed in a clean and dry alumina crucible. Then the alumina crucible was sealed in a tantalum tube with Ar at a pressure of about 0.7 bar, which restrained the volatilization of Sn during the heating process. Finally, the tantalum tube was sealed in a quartz tube with an Ar environment at 2 mbar. The quartz tube was placed in a furnace and heated to $1,150^\circ\text{C}$ from room temperature, kept at $1,150^\circ\text{C}$ for 48 h, cooled to 910°C in 6 h and then cooled to 800°C at a rate of 1.5°C h^{-1} . The quartz tube was moved quickly into the centrifuge to remove the excess Sn flux from Fe_3Sn_2 single crystals. The shape of a single crystal is hexagonal with shiny surfaces.

Transport and magnetic measurement. The longitudinal resistivity ρ_{xx} and interlayer (along the c axis) resistivity ρ_c were measured using a standard four-probe method on a Quantum Design physical properties measurement system. The magnetic moment was measured by using a Quantum Design magnetic property measurement system.

Extended Data Fig. 1 shows the longitudinal resistivity, ρ_{xx} , measured for temperatures from 380 K to 5 K. The residual resistivity ratio, defined as $\rho_{xx}(300\text{ K})/\rho_{xx}(5\text{ K})$, was calculated to be 58.4. This large number attests to the high quality of our single crystal studied here.

Extended Data Fig. 2 compares the in-plane and out-of-plane magnetization curves measured at 5 K. Although both curves show a saturated value close to $2\mu_B$ per Fe atom, the in-plane magnetization saturates much sooner, suggesting that the easy magnetization axis lies within the a – b plane. This is in agreement with our vector-field STM measurements, which showed that the spontaneous magnetization is along the a axis.

Extended Data Fig. 3 shows the evolution of the interlayer resistivity as a function of the azimuth angle of an in-plane field. Consistent with the vector-field STM study, it also reveals a two-fold-symmetric evolution with the azimuth angle of the field, demonstrating intrinsic electronic nematicity.

STS measurement on FeSn surface. As can be seen from Extended Data Fig. 4, strong interferences arise from native surface defects as either dark or bright spots in the topographic image. The bright spots are probably Sn adatoms and the dark spots are probably either Sn or Fe vacancies (see Extended Data Fig. 5 for details regarding these assignments), both of which can be generated by incomplete cleaving between adjacent FeSn and Sn layers. The Sn adatoms and Sn or Fe vacancies are originally non-magnetic scatterers and can cause potential scattering. It cannot be completely ruled out that there is an induced local moment around these scatterers in this correlated material.

Topological interpretation of the magnetization-controlled QPI symmetry. Figure 3 demonstrates a magnetization-direction-governed QPI symmetry (scattering vector q_1) associated with a band bottom of a massive Dirac dispersion. This set of data suggests that the electronic structure couples with the magnetization direction owing to strong SOC. Here we provide a more detailed explanation of this interpretation at a phenomenological level. In a strong SOC-coupled material, the massive Dirac band always has a spin–momentum-locked texture. We assume the electronic states near this massive Dirac band bottom¹⁸ to have a helical spin–momentum texture. As shown in Extended Data Fig. 6a, in the case of c -axis magnetization, all spins have a c -axis component and non-magnetic backscattering can happen in all in-plane directions. In the case of in-plane magnetization, the band will have a perpendicular momentum shift so as to gain a non-zero net spin component along the magnetization direction (Extended Data Fig. 6b, c; similar to the Edelstein effect^{10,31}). We note that the spins are still locked perpendicular to the original momentum centre (black circle). Nonmagnetic backscattering is then allowed along the magnetization direction but is forbidden along the perpendicular direction owing to spin reversal. Consistent with this picture, the scattering probability can be calculated under a $\mathbf{k} \cdot \mathbf{p}$ model, which manifests as a two-fold nodal function $I(\theta) \propto m^2 \cos^2 \theta / (v^2 k_F^2)$, where m is the in-plane magnetization strength, v is the band velocity, k_F is the Fermi momentum and θ is defined as the in-plane angle with respect to the magnetization direction.

Although the detailed spin-momentum texture in real materials can be more complex, such a mechanism for magnetization-governed selective scattering of massive Dirac fermions captures the essential aspect of our observation (including QPI symmetry and the predominantly nonmagnetic nature of the scattering source/agent) and offers consistency between angle-resolved photoemission spectroscopy (ARPES) and STM in explaining the data in Fig. 3a–c.

Discussion of possible effects of skyrmions on the magnetization-induced energy shift. Recently, it has been shown that magnetic skyrmions in magnetic thin films can induce a field-dependent energy shift²⁶. Spontaneous magnetic skyrmion excitations in Fe_3Sn_2 have also been reported²⁰. Therefore, it is necessary to discuss the possibility of such an interpretation of the magnetization-induced energy shift (Fig. 2).

The first point of consideration is that the spectra in Fig. 2 that exhibit the energy shift are all taken at the same positions, far from any defects. The STS images in Extended Data Fig. 4 show that the electronic structure is homogeneous far from defects, as seen in our samples, and there is no detectable spectroscopic evidence for skyrmions in the STM and STS results. Moreover, unlike non-centrosymmetric magnets and multilayer films, where the Dzyaloshinskii–Moriya interaction is crucial for the formation of skyrmions, our Fe_3Sn_2 is a centrosymmetric magnet, so the Dzyaloshinskii–Moriya interaction is not directly relevant. In such a system, ideally the magnetic dipole–dipole interaction and the uniaxial magnetic anisotropy ($K_{u\perp}$) have roles during the formation of skyrmions^{32,33}. Previous neutron scattering studies¹⁹ and magnetic measurements^{14,20} show that with decreasing temperature, the easy magnetization axis of Fe_3Sn_2 gradually changes from the c axis to the a – b plane. This translates to $K_{u\perp}$ being reduced as the temperature is lowered. A micromagnetic simulation based on these data suggests that skyrmions can only appear for moderate magnitudes of $K_{u\perp}$ and will thus disappear below a certain temperature²⁰. Consistent with this simulation and the experiments reported in ref. ²⁰, our in situ Lorentz transmission electron microscopy images at different temperatures (Extended Data Fig. 7) clearly demonstrate that the skyrmions vanish as the temperatures drop to below 130 K in Fe_3Sn_2 . Therefore, it is unlikely that at the working temperature of our STM and STS measurements spontaneous skyrmions can have any role, in accordance with our Lorentz transmission electron microscopy images.

Lastly, we discuss the possibility of field-induced skyrmions at the operating temperature of our STM and STS measurements and their effect on the energy shift. It is important to note that if the magnetic field is high enough, the magnetic domains (magnetic stripes, skyrmions and vortices) will evolve into ferromagnetic states and there is no peak shift with respect to the ferromagnetic background²⁸. As the magnetization saturates around 1 T for $\mathbf{B} \parallel c$ and at 0.5 T for $\mathbf{B} \parallel a$ – b for Fe_3Sn_2 (Extended Data Fig. 2), skyrmions will disappear in such field conditions. Accordingly, the field-induced skyrmions will exhibit a non-monotonic behaviour as the field increases, in contrast to the monotonic saturation behaviour shown in Fig. 2. Therefore, it is unlikely that the field-induced energy shift that we observed is related to skyrmions.

Comparison with the non-atomic cleavage surface of Mn_3Sn single crystals. Although the weaker interlayer bonding between the honeycomb Sn layer and the FeSn bilayer allows the preferred cleaving planes to exist in Fe_3Sn_2 , the Mn_3Sn bonds along the c axis and within the a – b plane are almost equivalent (similar to the situation within the FeSn bilayer), leaving no natural atomic cleaving plane. This is also suggested by the three-dimensional crystal shape of Mn_3Sn and two-dimensional crystal shape of Fe_3Sn_2 , as shown in Extended Data Fig. 8a, b. We have also successfully cleaved Mn_3Sn single crystals and imaged flat surfaces. However, they all exhibit non-atomic structures, as shown in Extended Data Fig. 8c, d, for example. These images are in sharp contrast to those measured on Fe_3Sn_2 , indicating that Fe_3Sn_2 is ideal for performing state-of-the-art STM studies in the transition-metal stannide family.

Theoretical discussion of the magnetization-direction-dependent Dirac gap. The model of massive Dirac fermions in a ferromagnetic kagome metal describes the Fe_3Sn_2 band structure obtained by ARPES rather well¹⁸. However, this description depends on the direction of the net magnetization. For simplicity, we can first ignore the interlayer coupling, which does not change the conclusion. The effective Hamiltonian can be written as

$$H = H_k + H_{\text{SOI}} + H_m$$

where H_k represents the tight-binding model of the kagome lattice with nearest-neighbour hopping, t :

$$H_k = \sum_{ij} t c_i^\dagger c_j$$

Here c_j (c_i^\dagger) is the electron annihilation (creation) operator in the spinor notation. H_{SOI} is the Kane–Mele-type spin–orbit interaction (SOI):

$$H_{\text{SOI}} = i \sum_{ij} \lambda_{ij} (c_i^\dagger s_z c_j)$$

where λ is the SOI amplitude, s_z is the spin Pauli matrix and $v_{ij} = 2(\mathbf{d}_i \times \mathbf{d}_j) \cdot \mathbf{z} / \sqrt{3}$, with \mathbf{d}_i and \mathbf{d}_j denoting the unit vectors along the two bonds that the electron traverses from site i to site j on the kagome lattice, as shown in Extended Data Fig. 9. We note that the Kane–Mele SOI only contains spin s_z without spin-flipping; thus, the z component of the spin is conserved.

H_m is the double-exchange coupling, J_H , between the ferromagnetic moment \mathbf{m}_i and the conduction electrons³:

$$H_m = -J_H \sum_i c_i^\dagger \mathbf{s} \cdot \mathbf{m}_i c_i$$

Clearly, when J_H or \mathbf{m}_i is zero, H generates the massive Dirac fermions as discussed in ref.¹⁸. Because Fe_3Sn_2 is ferromagnetic with a large ferromagnetic moment, the energy scale of $J_H \mathbf{m}_i$ can be large relative to the SOI λ . The magnetization \mathbf{m}_i splits the spin-degenerate bands into the spin majority and minority bands, with their separation controlled by J_H and \mathbf{m}_i .

By diagonalizing the Hamiltonian H , we find that the Dirac mass gap depends on the direction of \mathbf{m}_i , as shown in Extended Data Fig. 10. When the magnetization is $\mathbf{m}_i = m\hat{z}$, that is, along the z direction (M_c), H_{SOI} causes scattering of the electrons within the spin minority and majority sectors individually. This opens the Dirac mass gap, as shown in Extended Data Fig. 10a by the red lines. On the other hand, if \mathbf{m}_i is in the x - y plane (M_{a-b}), H_{SOI} scatters the electrons between the well separated spin minority and majority sectors. Thus, H_{SOI} cannot open the Dirac mass gap, as shown in Extended Data Fig. 10a by the blue lines. Experimentally, the magnetization direction of the ferromagnetic order in Fe_3Sn_2 is along the \mathbf{a} direction in the x - y plane. In this sense, H cannot explain the gap opening observed in ARPES. The same conclusion is reached in the case when interlayer coupling is included, where the band structure is shown in Extended Data Fig. 10b.

Because the Kane–Mele SOI cannot open the Dirac mass gap when the polarizing ferromagnetic moment lies in the a - b plane, the description of the electronic structure in Fe_3Sn_2 must go beyond the Kane–Mele approach. There are at least two possibilities, as discussed in the main text. The first one is to consider other types of or more complete spin–orbit coupling terms based on the full crystal symmetry. Such a microscopic description is difficult at the present time. Another possibility, which we briefly describe here, is that the ferromagnetic order in Fe_3Sn_2 is non-coplanar with a non-zero spin chirality³. In the large J_H limit, the conduction electron spin is forced to align with the spin direction at each site. Then, the hopping terms acquire a Peierls phase resulting from rotating the electron spin to the local spin direction. This gives rise to a spin Berry phase Ω_{ijk} .

As shown in Extended Data Fig. 11a, the spin chirality $\chi_{ijk} = \mathbf{S}_i \cdot (\mathbf{S}_j \times \mathbf{S}_k)$ corresponds to the spin Berry phase Ω_{ijk} , which is equal to half the solid angle formed by the spins \mathbf{S}_i , \mathbf{S}_j and \mathbf{S}_k . An electron hopping around this spin configuration acquires a gauge flux Ω and opens the Dirac gap on the kagome lattice³. Interestingly, this

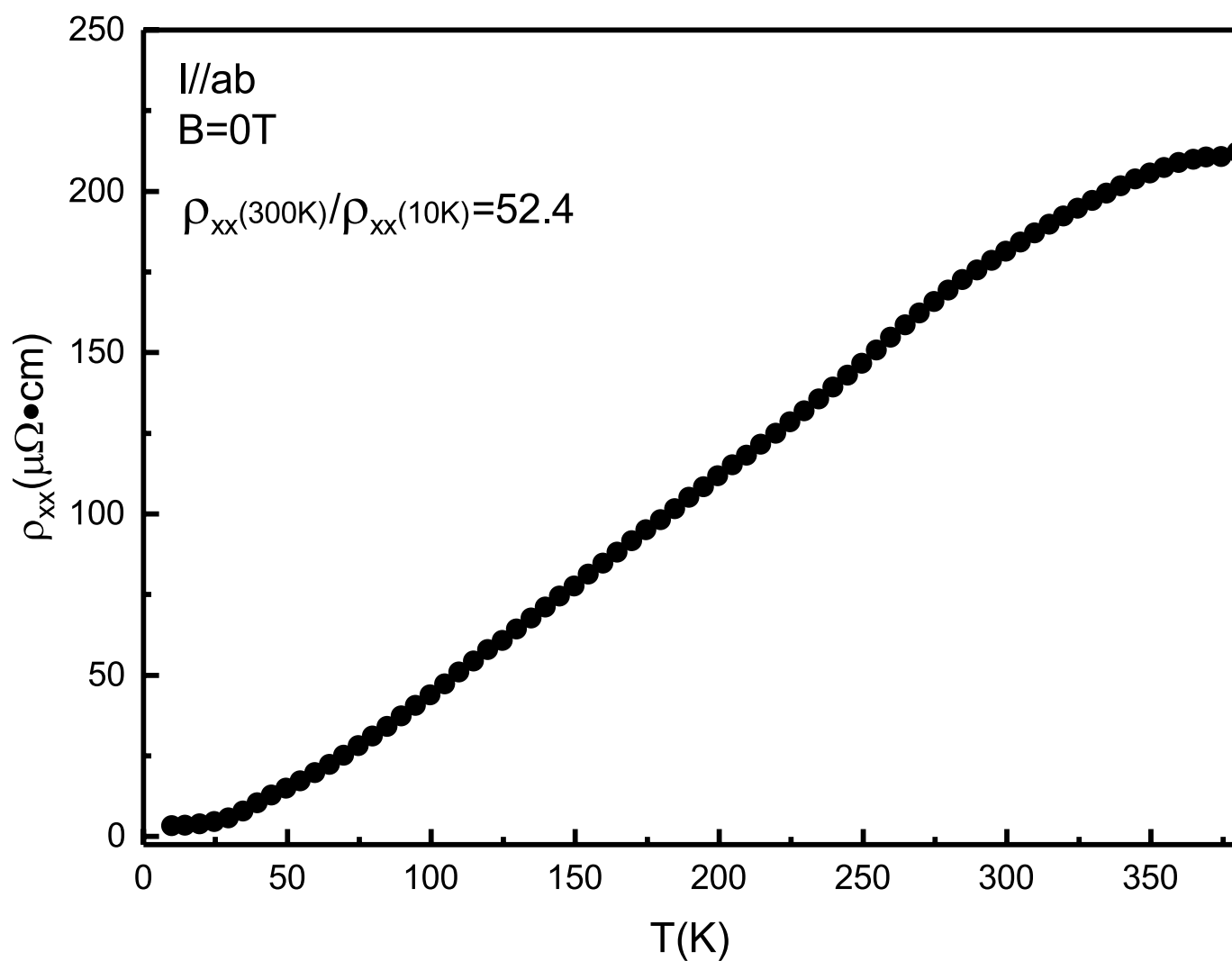
mass gap is independent of the direction of the net magnetization. By including both the Kane–Mele SOI and the spin Berry phase, the effective model in the local spin basis in the spin minority sector can be written as

$$H_{\text{eff}} = \sum_{ij} t e^{i\Omega/3} f_i^\dagger f_j + i\lambda_{ij}(\theta) f_i^\dagger f_j$$

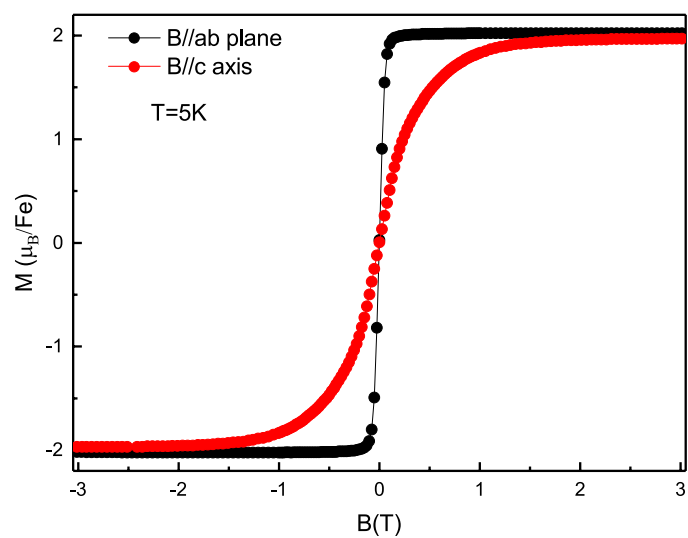
for $t \ll \lambda$, where f_i is the fermion operator in the rotated basis and the effective SOI strength $\lambda_{ij}(\theta)$ depends on the net magnetization direction measured by the angle θ away from the plane. When the magnetization lies in the plane, $\lambda_{ij}(\theta) = 0$, while when it is along the z direction, $\lambda_{ij}(\theta) \propto \lambda$. As a result, the spin Berry phase from a non-coplanar ferromagnetic order opens the Dirac gap while λ does not contribute to the gap opening in the absence of an external field. When an external magnetic field rotates the saturated ferromagnetic order to the z direction, the SOI becomes important and interferes with the spin Berry phase Ω . As shown in Extended Data Fig. 11c, when the magnetization lies in the x - y plane (M_{a-b}), Ω dominates and opens the Dirac gap. Furthermore, the Dirac gap can become smaller ($\lambda\Omega > 0$) when the magnetization is along the z direction (M_c), consistent with our interpretation of the experimental data. It is also possible that the Dirac gap for in-plane magnetization arises from purely orbital mechanisms (such as the effects of orbital hybridization discussed in ref.²⁵ and staggered orbital order^{34,35}), which competes with the Kane–Mele SOC when magnetization is rotated to lie along the c axis³⁶. However, given the existence of large anomalous Hall and related Berry-curvature effects established by transport measurements in this material, we adopt a Berry-curvature interpretation, which can then consistently describe both transport and spectroscopic data.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

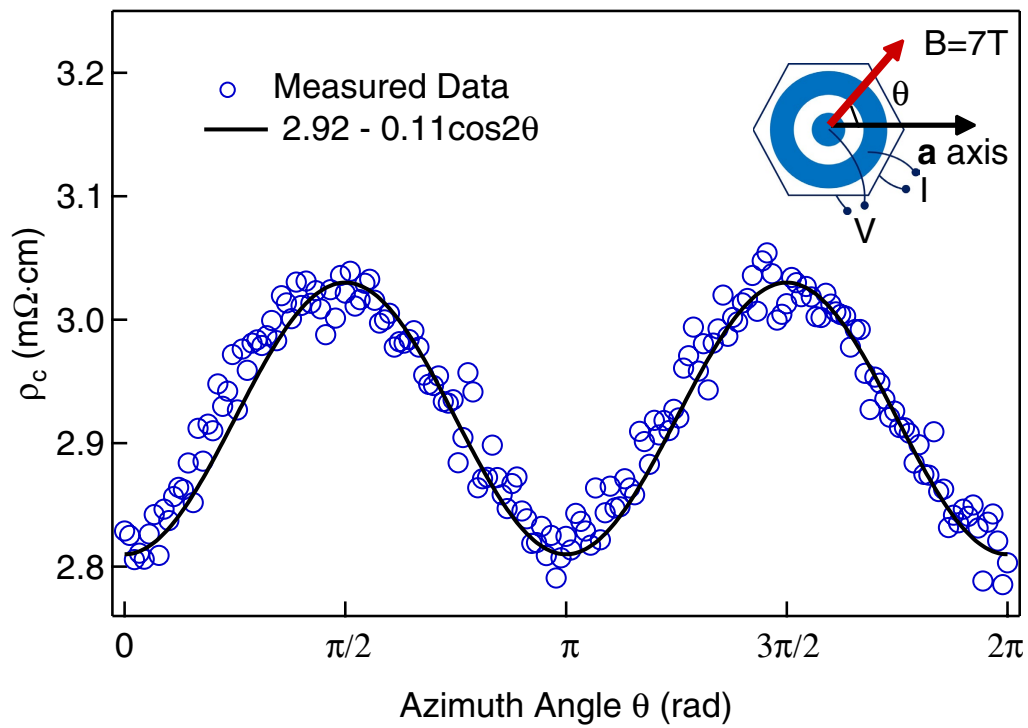
31. Edelstein, V. M. Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems. *Solid State Commun.* **73**, 233–235 (1990).
32. Seki, S. & Mochizuki, M. *Skyrmions in Magnetic Materials* 22–26 (Springer, Cham, 2016).
33. Montoya, S. A. et al. Tailoring magnetic energies to form dipole skyrmions and skyrmion lattices. *Phys. Rev. B* **95**, 024415 (2017).
34. Haldane, F. D. M. Model for a quantum Hall effect without Landau levels: condensed matter realization of the “parity anomaly”. *Phys. Rev. Lett.* **61**, 2015–2018 (1988).
35. Hasan, M. Z. et al. topological insulators, helical topological superconductors and Weyl fermion semimetals. *Phys. Scr.* **T164**, 014001 (2015); corrigendum **T168**, 019501 (2016).
36. Xu, S. Y. et al. Hedgehog spin texture and Berry’s phase tuning in a magnetic topological insulator. *Nat. Phys.* **8**, 616–622 (2012).



Extended Data Fig. 1 | Longitudinal resistivity measurement. The longitudinal resistivity ρ_{xx} was measured from 380 K to 5 K at zero magnetic field. The current was applied along the a - b plane.

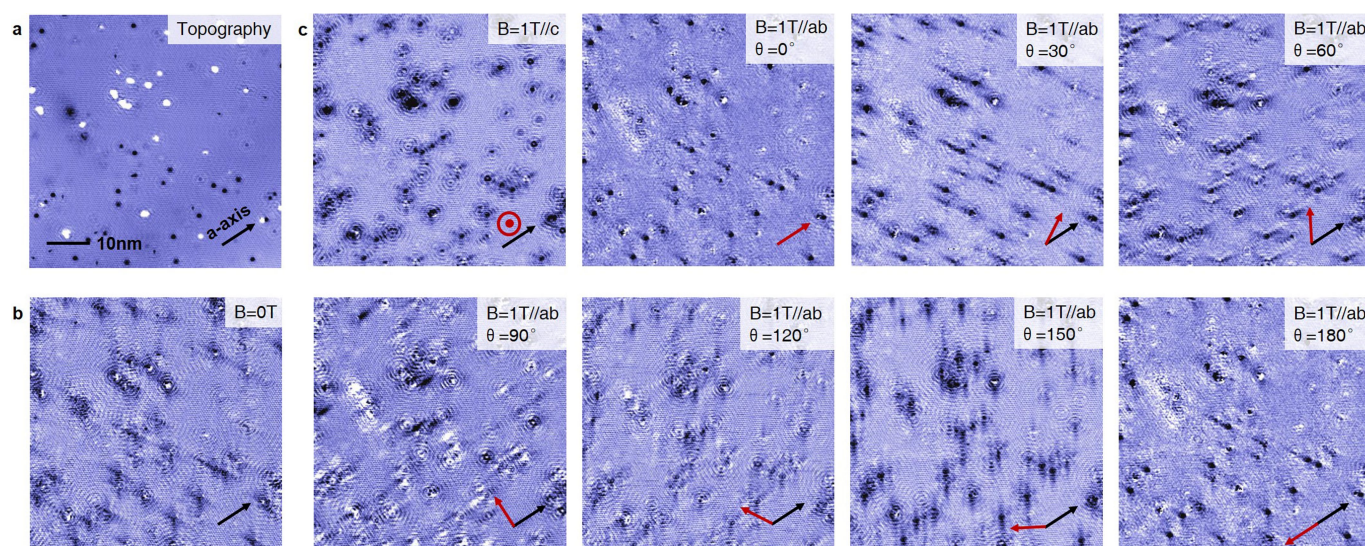


Extended Data Fig. 2 | Magnetization at 5 K in different magnetic field directions. The sharp saturation of the magnetization curve for the in-plane field suggests that the easy magnetization axis lies in the a - b plane.



Extended Data Fig. 3 | Interlayer resistivity evolution as a function of in-plane field azimuth angle. The black line is the fitting curve. The inset shows a schematic of the sample (black hexagon) with the concentric

electrical contacts used in such measurements (inner blue circle, voltage contacts; outer blue ring, current contacts).

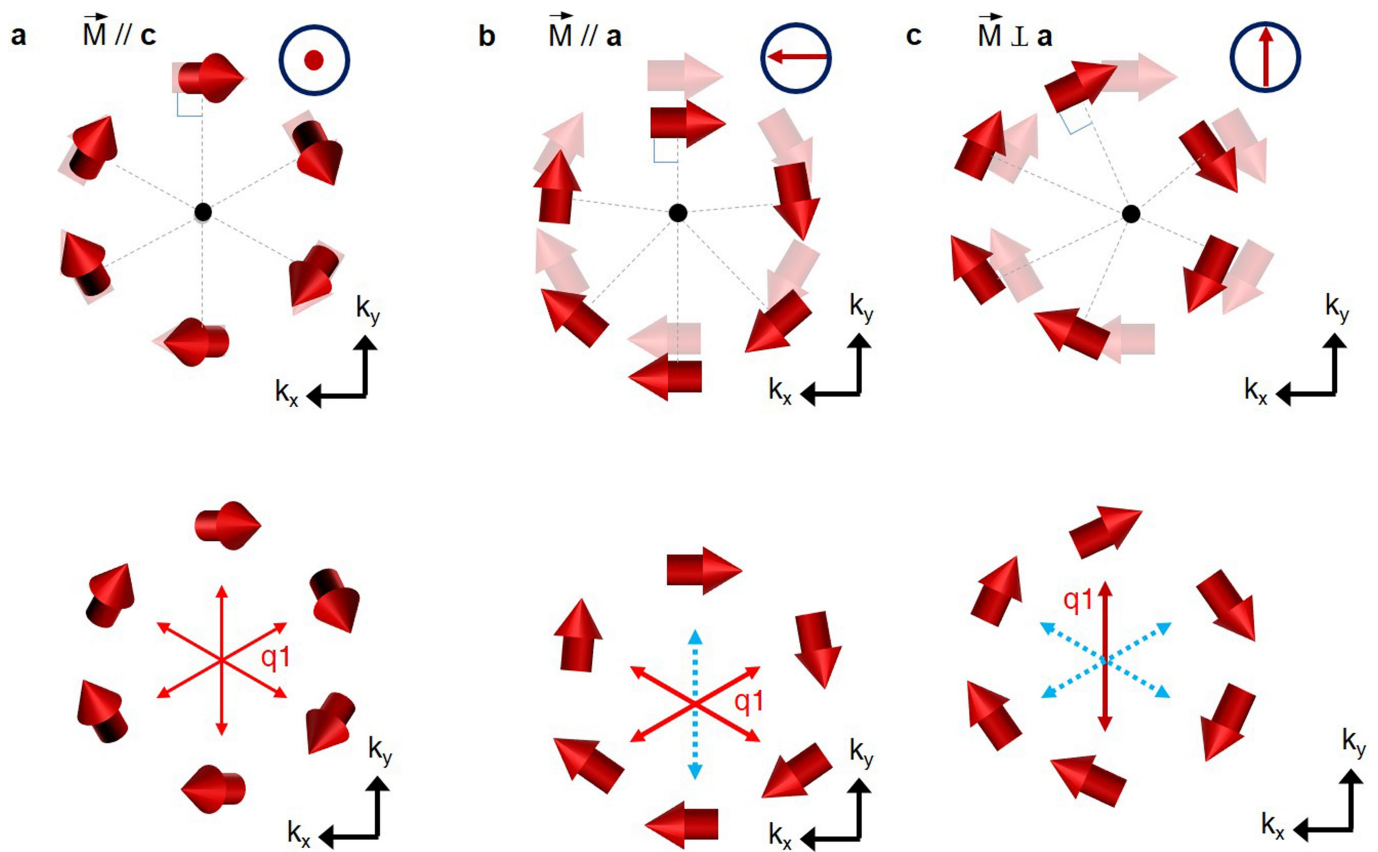


Extended Data Fig. 4 | Spectroscopic maps with vector-field conditions. **a**, Topography of a large FeSn surface. **b**, Differential conductance map taken in the area shown in **a** at the energy of the side peak. **c**, Differential conductance maps with various vector-field conditions taken in the area

shown in **a** at the energy of the side peak. The black arrow indicates the *a* axis and the red arrow represents the field vector. Their corresponding FFT images (known as QPI signals) are shown in Fig. 3.

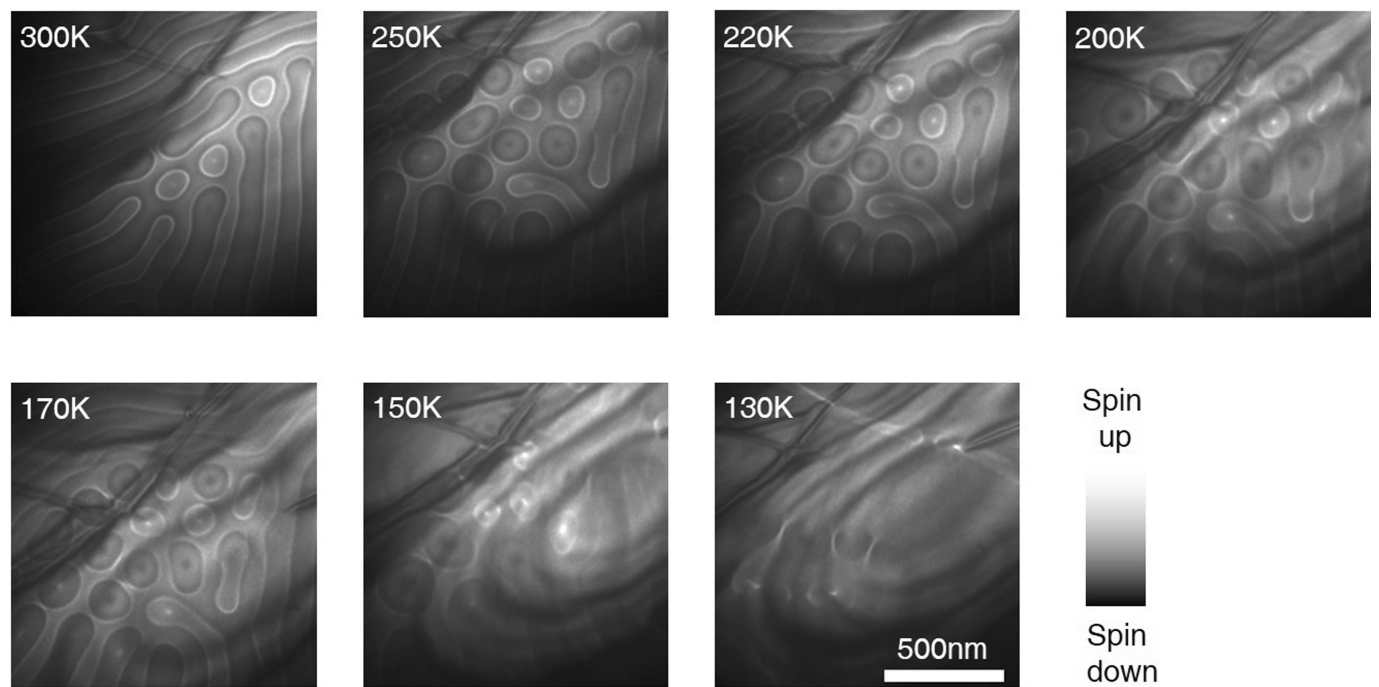


Extended Data Fig. 5 | Topographic image of a FeSn surface, showing two kinds of dark-spot defect. The centres of the defects are located at Fe and Sn sites, based on the atomic assignment from Fig. 1.



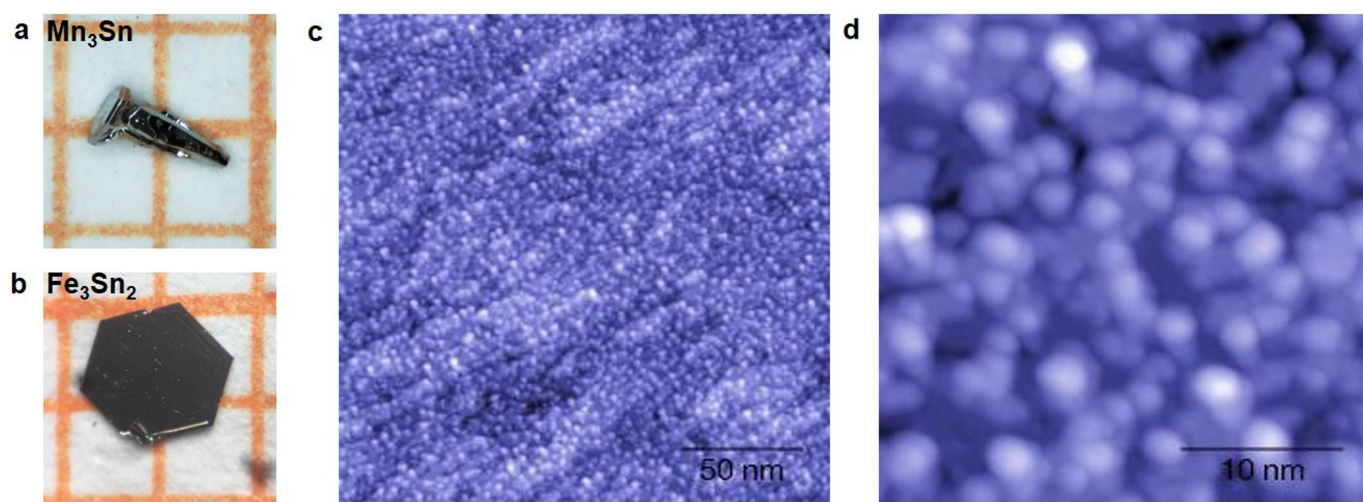
Extended Data Fig. 6 | Nonmagnetic scattering channels for helical spin-momentum texture with different magnetization directions.
a, The upper panel shows that spins cant towards the c axis to generate a net c -axis magnetization. The lower panel shows that backscattering is allowed in all directions. **b, c,** To generate a net in-plane magnetization, the momentum of the band shifts perpendicular to the in-plane direction and the spins reorganize their directions (similar to the Edelstein effect³¹),

as shown in the upper panels (the faded red arrows represent helical spin texture without magnetization for reference). The spins are still locked perpendicular to the original momentum centre³¹ (black circle). The lower panels show that backscattering is allowed only in the magnetization direction and forbidden in the perpendicular direction owing to spin reversal.

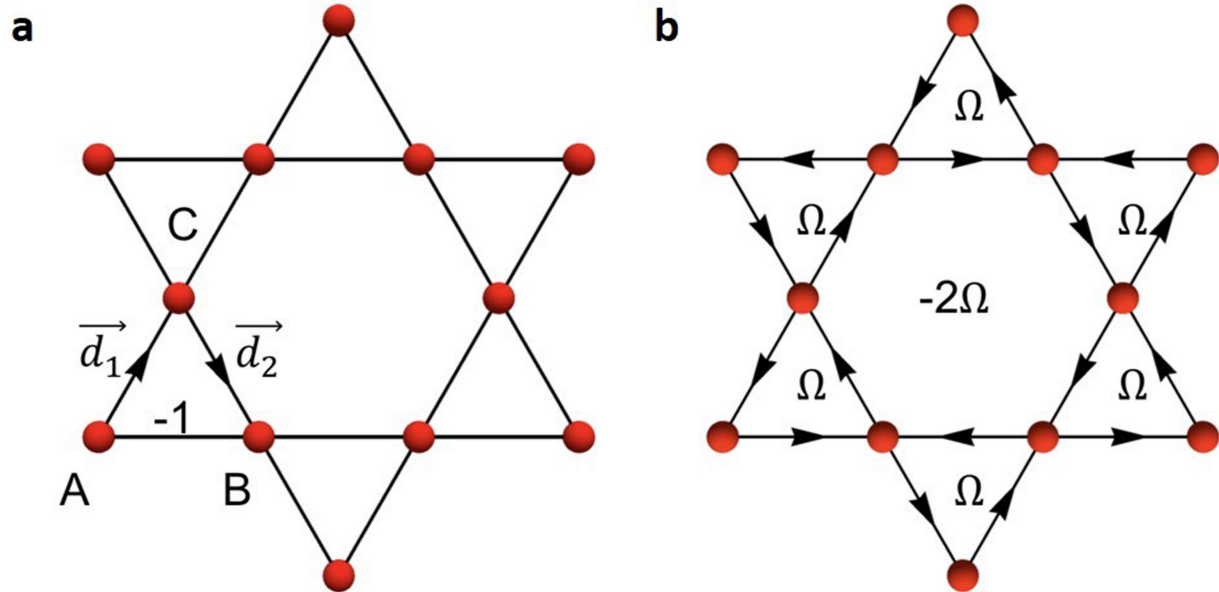


Extended Data Fig. 7 | Lorentz transmission electron microscopy images of the magnetic domain structures taken at different temperatures after turning off the external 0.7-T magnetic field applied along the c axis. The measurement is taken in the a - b plane in these

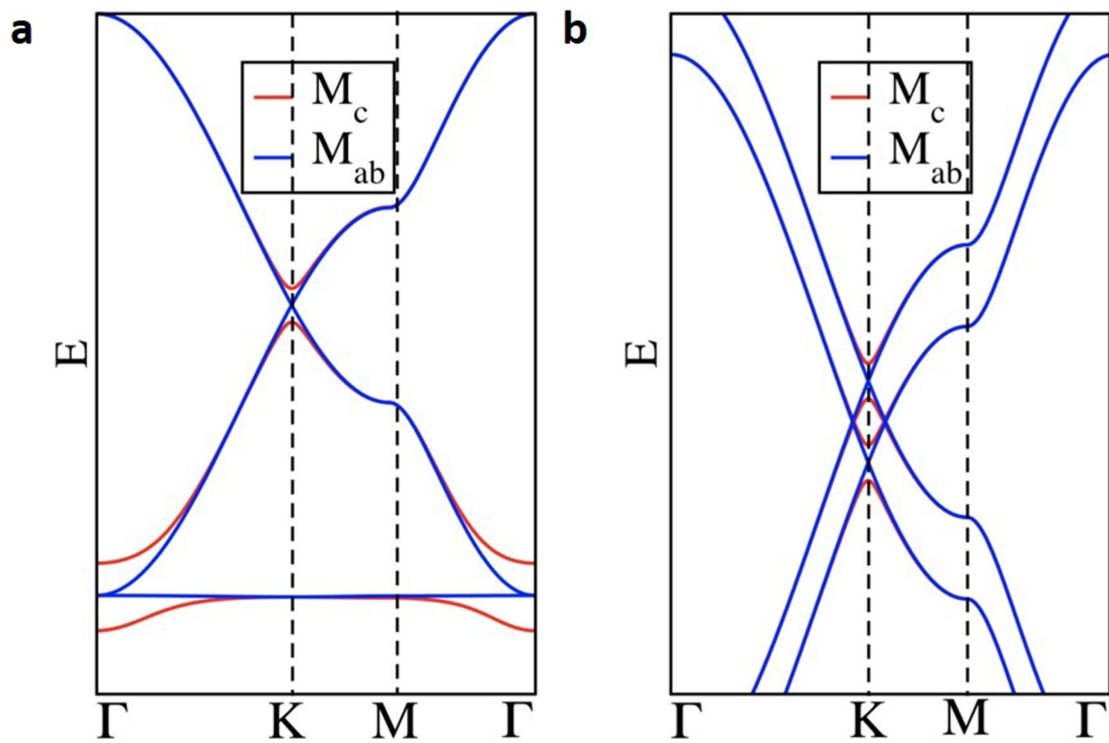
images and the different shades indicate the different orientation of the spin relative to the a - b plane. The skyrmions and stripe domains gradually disappear when the temperature drops to 130 K.



Extended Data Fig. 8 | STM results on Mn_3Sn . **a, b**, Single crystals of Mn_3Sn and Fe_3Sn_2 , respectively. **c, d**, Typical STM images of the cleavage surface of Mn_3Sn showing no clear atomic lattice structure.

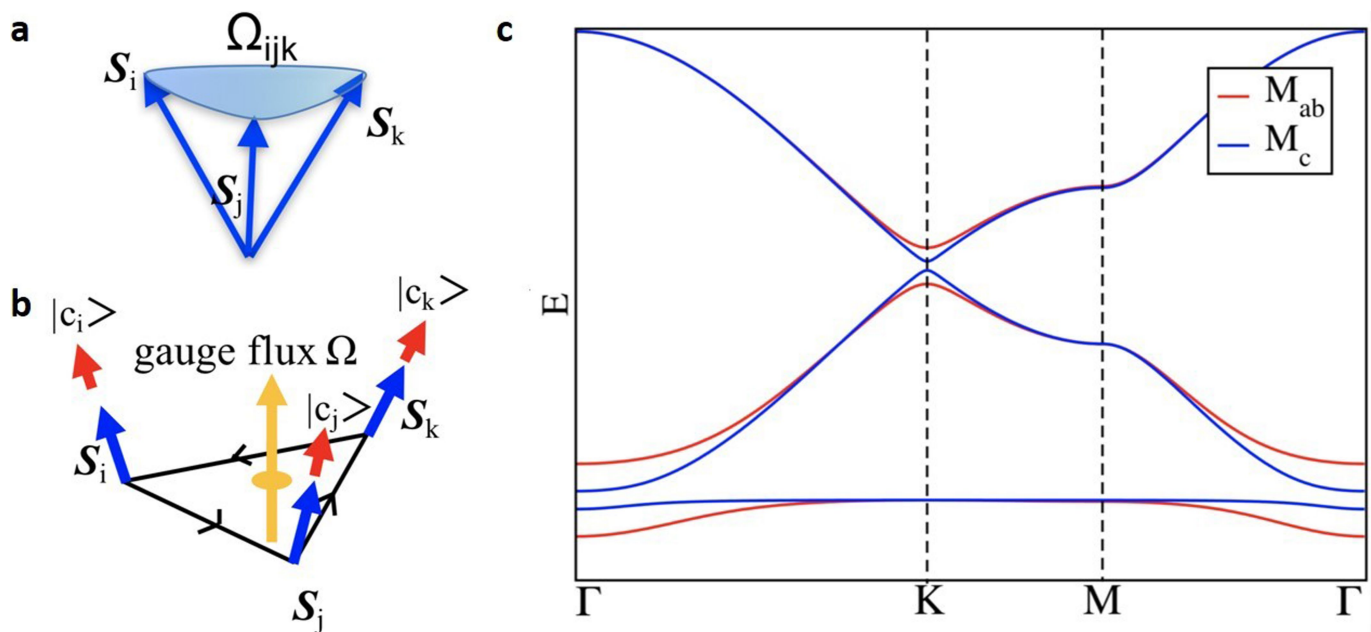


Extended Data Fig. 9 | Kagome lattice model. **a**, Kagome lattice with Kane-Mele SOI. Here, $v_{AB} = 2(\vec{d}_1 \times \vec{d}_2) \cdot \hat{z} / \sqrt{3} = -1$, where \vec{d}_1 is the unit vector from A to C and \vec{d}_2 is the unit vector from C to B. **b**, Flux Ω enclosed by the fundamental plaquette formed by sites A, B and C on the kagome lattice.



Extended Data Fig. 10 | Band dispersion from the kagome model. a, Red lines are the band structure for M_i along the z direction (M_c) with a Dirac mass generated by λ . Blue lines are the band structure for M_i in the x - y

plane (M_{ab}), showing a vanishing Dirac mass gap even in the presence of λ . **b,** Similar to **a** but including bilayer splitting due to interlayer coupling. Only the spin minority bands are shown.



Extended Data Fig. 11 | Effect of spin Berry phase. **a**, The spin Berry phase Ω_{ijk} is equal to half the solid angle formed by spins S_i , S_j and S_k . **b**, An electron hopping in the chiral spin background is equivalent to

hopping in the presence of a gauge flux Ω . **c**, Band structures in the presence of both λ and chiral flux Ω , when the magnetic field and the magnetization are along the x - y plane (M_{ab}) and the z direction (M_c).

Ferroelectric polymers exhibiting behaviour reminiscent of a morphotropic phase boundary

Yang Liu¹, Haibibu Aziguli¹, Bing Zhang², Wenhan Xu¹, Wenchang Lu², J. Bernholc² & Qing Wang^{1*}

Piezoelectricity—the direct interconversion between mechanical and electrical energies—is usually remarkably enhanced at the morphotropic phase boundary of ferroelectric materials^{1–4}, which marks a transition region in the phase diagram of piezoelectric materials and bridges two competing phases with distinct symmetries^{1,5}. Such enhancement has enabled the recent development of various lead and lead-free piezoelectric perovskites with outstanding piezoelectric properties for use in actuators, transducers, sensors and energy-harvesting applications^{5–8}. However, the morphotropic phase boundary has never been observed in organic materials, and the absence of effective approaches to improving the intrinsic piezoelectric responses of polymers^{9,10} considerably hampers their application to flexible, wearable and biocompatible devices. Here we report stereochemically induced behaviour in ferroelectric poly(vinylidene fluoride-co-trifluoroethylene) (P(VDF-TrFE)) copolymers, which is similar to that observed at morphotropic phase boundaries in perovskites. We reveal that compositionally tailored tacticity (the stereochemical arrangement of chiral centres related to the TrFE monomers^{11,12}) can lead to intramolecular order-to-disorder evolution in the crystalline phase and thus to an intermediate transition region that is reminiscent of the morphotropic phase boundary, where competing ferroelectric and relaxor properties appear simultaneously. Our first-principles calculations confirm the crucial role of chain tacticity in driving the formation of this transition region via structural competition between the *trans*-planar and 3/1-helical phases. We show that the P(VDF-TrFE) copolymer with the morphotropic composition exhibits a longitudinal piezoelectric coefficient of -63.5 picocoulombs per newton, outperforming state-of-the-art piezoelectric polymers¹⁰. Given the flexibility in the molecular design and synthesis of organic ferroelectric materials, this work opens up the way for the development of scalable, high-performance piezoelectric polymers.

The morphotropic phase boundary (MPB) was first observed in lead zirconate titanate (PZT) more than half a century ago¹³ and has turned into the most sought-after concept in the field of piezoelectricity because of its inherent, substantially enhanced piezoelectric effect^{1–4}. The challenge in constructing an MPB mainly arises from the design of structural competition, which facilitates polarization rotation between two nearly energetically degenerate phases in the phase diagram^{1,5}. The MPB has been observed in only a few perovskites, such as PZT^{2,3}, Pb(Zn_{1/3}Nb_{2/3})O₃-PbTiO₃ (PZN-PT)^{6,14}, Pb(Mg_{1/3}Nb_{2/3})O₃-PbTiO₃ (PMN-PT)^{4,15} and other systems^{5,7,8}, which has stimulated unprecedented interest in both fundamental research and practical applications. However, this physical concept has never been demonstrated in organic piezoelectric materials, which hold great promise for flexible and biofriendly applications^{9,10}.

Here, we consider the ferroelectric semi-crystalline P(VDF-TrFE) copolymers because the incorporation of various stereoirregular TrFE monomers may lead to different regiodefects and tacticity in these copolymers^{11,12}. Owing to a strong dependence of the conformational

energy on the polymer microstructure¹¹, the ground-state conformation could be subsequently transformed from the all-*trans* to the 3/1-helix conformation ((TG)₃ or (T \bar{G})₃; T, *trans*, G, *gauche*) (Extended Data Fig. 1). We therefore speculate that the conformational competition at different chemical compositions may drive the formation of an intermediate region that separates the 3/1-helical and *trans*-planar phases in the P(VDF-TrFE) copolymers and that resembles the MPB in perovskites.

We synthesized P(VDF-TrFE) copolymers with VDF contents (C_{VDF}) ranging from 45 mol% to 80 mol%. The tacticity of the polymers was quantified using ¹⁹F nuclear magnetic resonance (NMR) spectroscopy, by analysing the characteristic peaks of isotactic (mm), syndiotactic (rr) and heterotactic (mr+rm) triads (Extended Data Fig. 2). A schematic of the stereochemical configurations of P(VDF-TrFE) is presented in Fig. 1a. As shown in Fig. 1b, the VDF-TrFE segment remains predominantly syndiotactic (tacticity above 0.70), with modest heterotactic (tacticity below 0.25) and negligible isotactic contents (tacticity below about 0.06) at various chemical compositions. On the other hand, as shown in Fig. 1c, there exists a critical VDF content of $C_{\text{VDF}} = 49$ mol%, where the most favourable sequence for the TrFE-TrFE segment is changed from syndiotactic to isotactic. Computations on poly(trifluoroethylene) (PTFE)¹¹ have suggested that the all-*trans* conformation is more energetically favourable for the syndiotactic TrFE-TrFE sequence, whereas the 3/1-helical conformation is energetically preferred for the isotactic TrFE-TrFE sequence. We therefore believe that the critical change of the tacticity distribution in the TrFE-TrFE segment at $C_{\text{VDF}} \approx 49$ mol% could cause competition between the *trans*-planar and 3/1-helical phases, and that a transition region similar to the MPB in perovskites occurs. In addition, we found that the content of H-H/T-T (H, head; T, tail) regiodefects calculated from ¹⁹F NMR is almost independent of the polymer composition (Extended Data Fig. 2). Therefore, the possibility of using regiodefects to induce phase evolution^{11,16,17} and MPB formation was excluded.

The copolymers were characterized by X-ray diffraction (XRD) to confirm the existence of MPB-like behaviour. The peak splittings were clearly seen in the copolymers with $C_{\text{VDF}} \leq 55$ mol% (Fig. 1d), which were assigned to the *trans*-planar phase at a high diffraction angle 2θ and the 3/1-helical phase at low 2θ ^{18,19}. Phase competition—rather than a simple two-phase mixture—is indicated by the continuous growth of the newly formed 3/1-helical phase peak at $C_{\text{VDF}} = 55$ mol% at the expense of the peak from the *trans*-planar phase, with decreasing C_{VDF} (Fig. 1d). The apparent shift of the peaks with the composition corresponds to a very large increase in the intermolecular lattice spacings (Fig. 1e). The changes of slope in the interchain spacing indicate the structural evolution from the *trans*-planar phase to the 3/1-helical phase via an intermediate two-phase coexistence region, thus suggesting the occurrence of an MPB-like transition region in the P(VDF-TrFE) copolymers.

Consistent with the NMR and XRD results, Fourier-transform infrared (FTIR) spectroscopy provides further evidence for intramolecular evolution from the all-*trans* to the 3/1-helical conformation (Extended

¹Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA, USA. ²Department of Physics, North Carolina State University, Raleigh, NC, USA. *e-mail: wang@matse.psu.edu

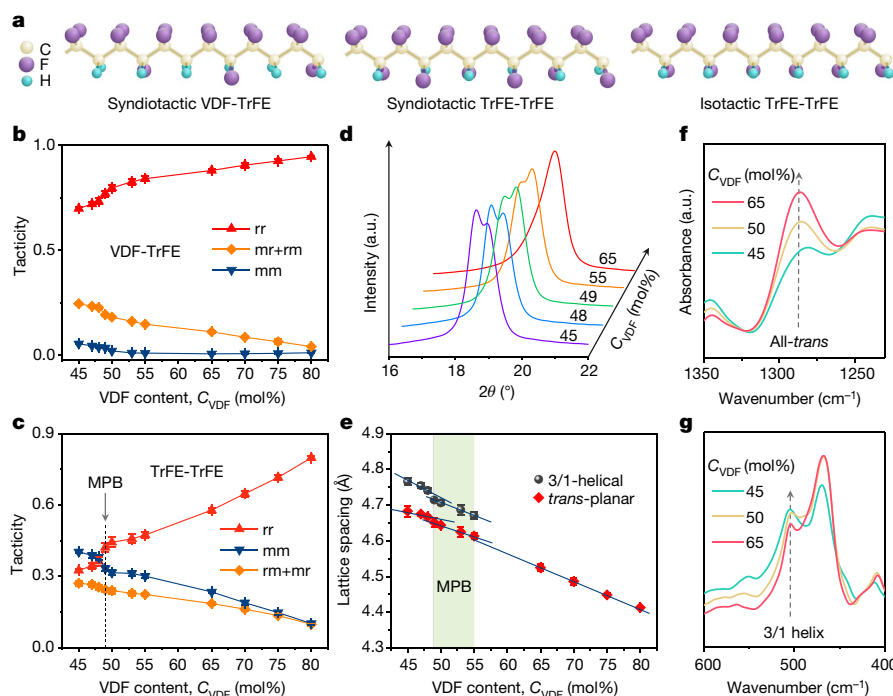


Fig. 1 | Structural evolution of P(VDF-TrFE) copolymers at various compositions. **a**, Sketch of chain tacticity in P(VDF-TrFE). **b**, Tacticity of the VDF-TrFE segment versus VDF content, C_{VDF} . The data were calculated by measuring the ratios of the integral intensities of the respective triad peaks in the -CHF- resonance region of the ^{19}F NMR spectra (Extended Data Fig. 2). Data are given for the isotactic (mm), syndiotactic (rr) and heterotactic (mr+rm) configurations. **c**, Tacticity of the TrFE-TrFE segment as a function of C_{VDF} . The TrFE-TrFE sequence evolves to be atactic-like, supported by a considerable increase in heterotactic (mr+rm) triads and a rapid development of isotactic (mm)

triads with decreasing C_{VDF} . The MPB-like transition occurs near the critical VDF content of 49 mol%, as indicated by the grey arrow. **d**, XRD θ - 2θ scans of copolymers with $C_{\text{VDF}} = 45$ –65 mol%. a.u., arbitrary units. **e**, Intermolecular lattice spacing versus C_{VDF} . The light-green-shaded area indicates the transition region, across which the structure changes abruptly. The solid lines in **b**, **c** and **e** are guides for the eyes and the error bars represent the standard deviation, obtained from at least three measurements using different samples. **f**, **g**, Infrared absorbance bands at around $1,290\text{ cm}^{-1}$ (**f**) and 507 cm^{-1} (**g**). The dashed arrows mark the characteristic peaks of the all-*trans* and 3/1-helical conformations.

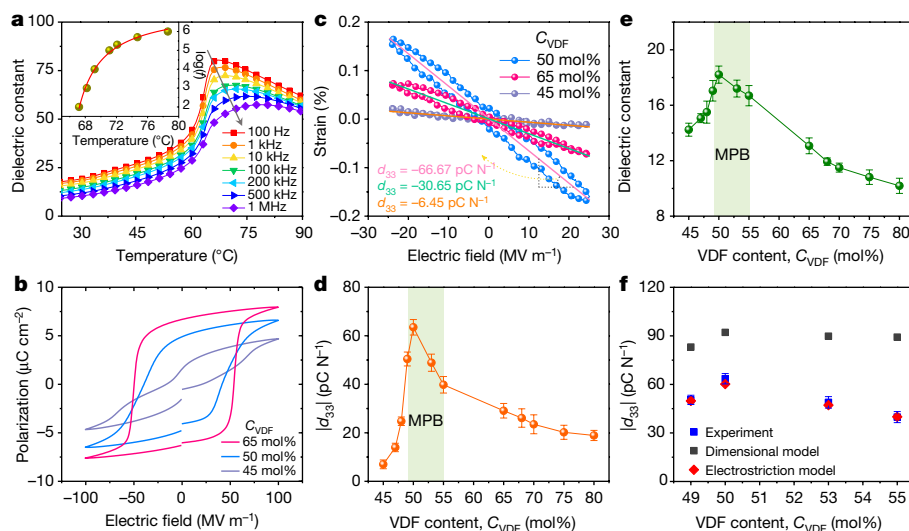


Fig. 2 | Dielectric, ferroelectric and piezoelectric properties of P(VDF-TrFE) copolymers. **a**, Temperature dependence of the dielectric constant of P(VDF-TrFE) ($C_{\text{VDF}} = 50\text{ mol}\%$), showing relaxor behaviour. The grey arrow shows the dependence of the dielectric constant on the frequency of the a.c. electric field upon heating. The inset shows a fit of the measured dielectric constant (dark yellow circles) with the Vogel–Folcher law (red solid line), $\ln f = \ln f_0 - U/[k_B(T_{\text{max}} - T_f)]$, where f is the frequency, f_0 is the attempt frequency, U is the activation energy, k_B is the Boltzmann constant and T_f is the freezing temperature. The fitting yields $U = 3.38 \times 10^{-3}\text{ eV}$, $f_0 = 1.91 \times 10^7\text{ Hz}$ and $T_f = 64.02^\circ\text{C}$. **b**, Polarization–electric field hysteresis loops, measured using a triangular-waveform a.c. electric field of 1 Hz at room temperature. **c**, Electric-field-induced strain, measured

using a 1-Hz triangular waveform of a bipolar electric field (25 MV m^{-1}) at room temperature. The longitudinal piezoelectric coefficient d_{33} was extracted from the slope of the strain–electric field curve. **d**, Magnitude of d_{33} as a function of C_{VDF} . **e**, Compositional dependence of the dielectric constant, measured at 1 kHz and room temperature. The light-green-shaded areas in **d** and **e** indicate the transition region, which shows substantially enhanced piezoelectricity and dielectric responses. **f**, Comparison of experimental d_{33} values with theoretical calculations from the dimensional and electrostriction models. Detailed descriptions of the models can be found in Methods. Error bars in **d**–**f** represent standard deviations obtained from at least three measurements using different samples.

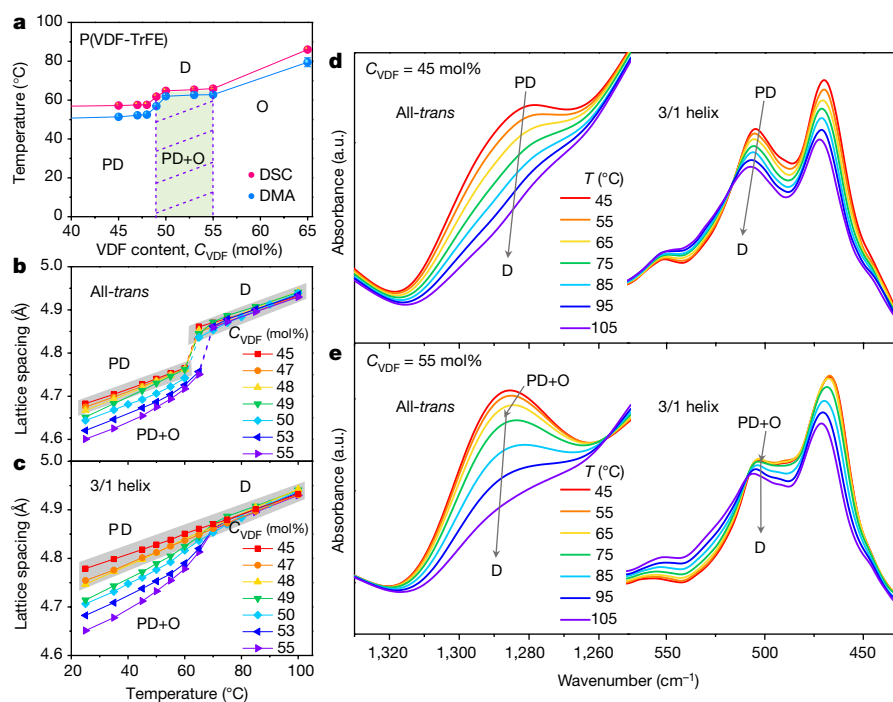


Fig. 3 | Phase diagram and transition near the intermediate region in P(VDF-TrFE) copolymers. **a**, Phase diagram around the transition region (light-green-shaded area between vertical dashed lines), obtained from DSC (red circles) and DMA (blue circles) measurements. The data show the transition to the paraelectric phase in the crystalline regions of P(VDF-TrFE). The labels ‘O’, ‘D’ and ‘PD’ denote the regions of the ordered, disordered and pseudo-disordered phases, respectively (Extended Data Fig. 1). Error bars (typically smaller than the symbols) represent standard deviations obtained from at least three measurements using different samples. **b**, **c**, Intermolecular lattice spacings of the *trans*-planar and 3/1-helical phases upon heating versus the temperature. The low-temperature phase (PD, PD+O) eventually transforms into the paraelectric phase (D) with nearly the same intermolecular lattice spacing for polymers with different VDF content. The grey-shaded regions

highlight the linear temperature dependence of interchain spacings. The dashed lines in **b** indicate that the doublet merges into a singlet (Extended Data Fig. 8). The solid lines in **a–c** are guides for the eyes. **d**, **e**, Temperature-dependent characteristic infrared absorbance bands of the all-*trans* conformation at about 1,290 cm^{−1} and for the 3/1-helical conformation at about 507 cm^{−1} of P(VDF-TrFE) copolymers with $C_{\text{VDF}} = 45$ mol% (**d**) and $C_{\text{VDF}} = 55$ mol% (**e**). The grey arrows qualitatively mark the boundaries between the low-temperature phase (PD+O, PD) and the high-temperature paraelectric phase (D), which are consistent with the results shown in **a–c**. At $C_{\text{VDF}} = 45$ mol%, the bands spread uniformly upon heating, indicating the absence of a phase transition. By contrast, at $C_{\text{VDF}} = 55$ mol%, two distinct regimes can be clearly resolved, suggesting the presence of an order–disorder phase transition.

Data Fig. 3). Specifically, the characteristic infrared absorbance at 1,290 cm^{−1} that corresponds to the all-*trans* conformation declines steadily with the decrease of C_{VDF} (Fig. 1f), accompanied by an increase of the peak at 507 cm^{−1}, which is assigned to the 3/1-helical conformation (Fig. 1g). Additionally, our analyses of the latent heat, measured using differential scanning calorimetry (DSC), and the loss tangent, obtained from dynamic mechanical analysis (DMA), substantiate the existence of an intermediate transition region (Extended Data Fig. 4).

This MPB-like behaviour is unambiguously supported by the electrical characterization of the P(VDF-TrFE) copolymers. Interestingly, relaxor ferroelectric properties were found in the P(VDF-TrFE) copolymers with $C_{\text{VDF}} \leq 55$ mol% (Extended Data Fig. 5). Figure 2a presents the typical frequency dependence of the dielectric peaks at the maximum temperature T_{max} in P(VDF-TrFE) with $C_{\text{VDF}} = 50$ mol%. This dependence is indicative of a relaxor ferroelectric and can be well described by the Vogel–Folcher law (see inset of Fig. 2a). Our finding reveals that the relaxor behaviour is actually intrinsic to the P(VDF-TrFE) copolymers and not triggered by irradiation²⁰. More importantly, distinctive ferroelectric hysteresis is evident in the polarization–electric field (*P–E*) loops of the copolymers with $C_{\text{VDF}} \geq 49$ mol% (Fig. 2b, Extended Data Fig. 6), indicating the coexistence of normal ferroelectricity with relaxor ferroelectricity in the morphotropic range of the P(VDF-TrFE) copolymers, that is, for 49 mol% $\leq C_{\text{VDF}} \leq 55$ mol%. As shown in Fig. 2b and Extended Data Fig. 6, the *P–E* loops of the copolymers with $C_{\text{VDF}} < 49$ mol% are characterized by antiferroelectric-like hysteresis²¹ due to stabilization of the local ferroelectric distortion in relaxors.

The piezoelectric and dielectric properties of the copolymers were carefully evaluated (Extended Data Figs. 5, 7) because the MPB improves the electromechanical and dielectric responses of ferroelectrics^{1–4}. The longitudinal piezoelectric coefficient d_{33} was measured using a direct probe of the electric-field-induced strain^{22,23} (Fig. 2c). As summarized in Fig. 2d, $|d_{33}|$ displays a unique λ -shaped curve versus C_{VDF} . The maximum piezoelectricity, $|d_{33}| = -63.5 \pm 3.2$ pC N^{−1}, is achieved for the P(VDF-TrFE) copolymer with $C_{\text{VDF}} = 50$ mol% and is approximately double the d_{33} values of copolymers away from the morphotropic region (see Supplementary Table 2). Similarly, the dielectric constant is maximized at $C_{\text{VDF}} = 50$ mol% (Fig. 2e). We also compare our piezoelectric results with theoretical models (Fig. 2f, Extended Data Fig. 7), including the dimensional model, which is based on the deformation of the amorphous regions²⁴, and the electrostriction model, which describes dimension changes of the crystalline regions in normal ferroelectrics (see Methods). We find that the electrostriction model agrees well with our experimental data, whereas the dimensional model yields substantial deviations (Fig. 2f). These findings explicitly reveal that piezoelectricity originates from the electrostriction in the crystalline domains of P(VDF-TrFE) copolymers, corroborating our conclusion about the crystalline nature of MPB.

Figure 3a presents the phase diagram of the temperature versus the composition of the P(VDF-TrFE) copolymers around the MPB-like transition region, as determined by DSC and DMA, respectively. The diagram shows a nearly rectangular intermediate region (width of about 6 mol%) with two vertical phase boundaries on both sides, where the solid lines denote the transition to the paraelectric phase. To understand the molecular structures near the transition region, we performed

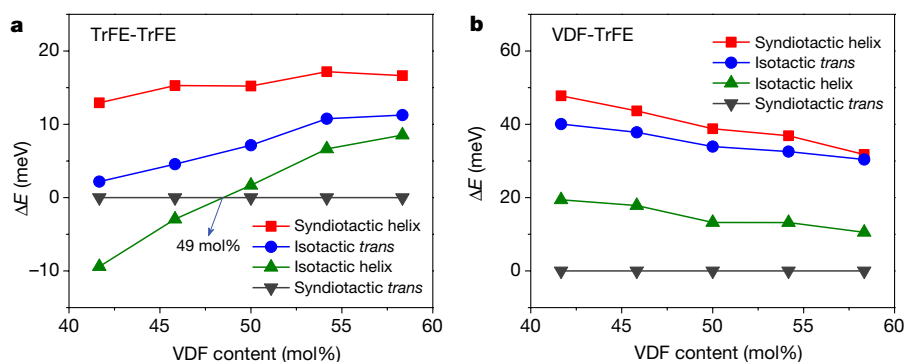


Fig. 4 | Influence of chain tacticity on the stability of the *trans*-planar and 3/1-helical phases in P(VDF-TrFE) copolymers near the transition region. **a**, Calculated energy difference ΔE between various phases with different tacticities and the syndiotactic *trans*-planar phase in the TrFE-TrFE segment. The steric interactions between the fluorine atoms may favour the syndiotactic *trans*-planar phase, which was chosen as the reference energy in our calculations. The monotonic decrease of ΔE with decreasing VDF content corresponds to a progressive stabilization of the

3/1-helical phase. Below the critical VDF content of 49 mol%, the 3/1-helical phase is more stable than the syndiotactic *trans*-planar phase, which agrees well with our experimental results (Figs. 1c, 3a). **b**, Calculated energy difference ΔE between various phases with different tacticities and the syndiotactic *trans*-planar phase in the VDF-TrFE segment. The ground state of the VDF-TrFE segment always corresponds to the syndiotactic *trans*-planar phase for the compositions of interest. The lines in **a** and **b** are guides for the eyes.

temperature-dependent XRD to describe the change in intermolecular lattice spacings (Extended Data Fig. 8) and variable-temperature FTIR to elucidate the intramolecular conformational evolution (Extended Data Fig. 9). No order–disorder phase transition was detected in the TrFE-rich region ($C_{\text{VDF}} < 49$ mol%), according to the linear temperature dependence of the intermolecular lattice spacings (Fig. 3b, c) and the uniform smearing of the infrared bands for the all-*trans* and 3/1-helical conformations upon heating (Fig. 3d). Consequently, the overall chain structure here resembles that of a disordered paraelectric phase that approximately comprises an irregular succession of the TG, T \bar{G} and TT groups^{18,19,25,26}. The copolymers with $49 \text{ mol}\% \leq C_{\text{VDF}} \leq 55 \text{ mol}\%$ present a typical order–disorder transition (Fig. 3b, c, e), suggesting that both the all-*trans* and the 3/1-helical conformations maintain a high degree of order. Within the intermediate transition region, the intramolecular disorder mainly develops in the 3/1-helical phase, whereas the fraction of the well ordered *trans*-planar phase decreases considerably with decreasing C_{VDF} . The changes in chain tacticity (Fig. 1c) not only stabilize the low-energy rotational isomers but also introduce concomitant conformational disorder (Extended Data Fig. 1). The evidence of this disorder being inherent to the morphotropic compositions comes from the experimental observation of a broad, arced and diffuse meridional reflection at about 2.31 \AA in the XRD spectra¹⁸, the broadening of the band at around 507 cm^{-1} in the FTIR data (Fig. 1g) and the emergence of relaxor behaviour in the dielectric spectra (Fig. 2a). Furthermore, the balance between two phases with similar energies is completely broken at $C_{\text{VDF}} < 49 \text{ mol}\%$, giving rise to the paraelectric-like disordered 3/1-helical phase. PTrFE was found to exhibit almost the same disordered conformation²⁷, supporting our conformational picture. We conclude that the compositional manipulation of the tacticity distribution leads to an order-to-pseudo-disorder evolution of the chain conformation (Extended Data Fig. 1). The right and left vertical boundaries shown in the phase diagram of Fig. 3a designate the appearance of pseudo-disorder and the disappearance of order, respectively.

We carried out first-principles calculations using density functional theory (DFT). Figure 4a, b shows the relative energy diagrams of the P(VDF-TrFE) copolymers near the transition region, where the syndiotactic *trans*-planar energy has been taken as a reference for the TrFE-TrFE and VDF-TrFE segments (Methods). We note that the contribution of the chain tacticity distribution was overlooked in most of the previous theoretical works. In this context, a recent DFT study reported the all-*trans* conformation to be the ground state of P(VDF-TrFE) copolymers²⁸. By contrast, we reveal that the energy minimum of P(VDF-TrFE) strongly depends on its chain tacticity. Specifically, with

decreasing C_{VDF} , the ground state of the TrFE-TrFE segment evolves from the syndiotactic *trans*-planar to the isotactic 3/1-helical phase at a critical value of $C_{\text{VDF}} \approx 49 \text{ mol}\%$ (Fig. 4a). By contrast, the most energetically favourable state of the VDF-TrFE segment is always the syndiotactic *trans*-planar phase (Fig. 4b). Our calculations validate the dramatic tacticity change observed experimentally in the TrFE-TrFE segment (Fig. 1c) and further confirm that this change is the origin of the MPB-like transition.

Our results necessitate future studies of the underlying symmetry near the transition region of the P(VDF-TrFE) copolymers (Extended Data Fig. 10). Given that the *trans*-planar lattice has been suggested to be orthorhombic ($Cm2m$ space group), like poly(vinylidene fluoride)²⁶, and the disordered paraelectric structure has been assumed to be hexagonal^{25,26,29} ($P6/mmm$ space group)²⁹ or pseudohexagonal^{18,19}, we expect symmetry breaking across the intermediate region revealed here. However, precise knowledge of intramolecular rotations in consecutive units in the polymer is lacking at present (Extended Data Fig. 1), which makes it challenging to define the periodicity along the chain axes. The discovery of MPB-like behaviour in these polymers offers opportunities for the development of high-performance organic piezoelectric materials and for the investigation of the MPB mechanism from the molecular perspective.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0550-z>.

Received: 1 May 2018; Accepted: 8 August 2018;

Published online 3 October 2018.

1. Fu, H. & Cohen, R. E. Polarization rotation mechanism for ultrahigh electromechanical response in single-crystal piezoelectrics. *Nature* **403**, 281–283 (2000).
2. Guo, R. et al. Origin of the high piezoelectric response in $\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$. *Phys. Rev. Lett.* **84**, 5423–5426 (2000).
3. Bellaiche, L., García, A. & Vanderbilt, D. Finite-temperature properties of $\text{Pb}(\text{Zr}_{1-x}\text{Ti}_x)\text{O}_3$ alloys from first principles. *Phys. Rev. Lett.* **84**, 5427–5430 (2000).
4. Kutnjak, Z., Petzelt, J. & Blinc, R. The giant electromechanical response in ferroelectric relaxors as a critical phenomenon. *Nature* **441**, 956–959 (2006).
5. Ahart, M. et al. Origin of morphotropic phase boundaries in ferroelectrics. *Nature* **451**, 545–548 (2008).
6. Park, S. E. & Shrout, T. R. Ultrahigh strain and piezoelectric behavior in relaxor based ferroelectric single crystals. *J. Appl. Phys.* **82**, 1804–1811 (1997).
7. Saito, Y. et al. Lead-free piezoceramics. *Nature* **432**, 84–87 (2004).
8. Zeches, R. J. et al. A strain-driven morphotropic phase boundary in BiFeO_3 . *Science* **326**, 977–980 (2009).
9. Lovinger, A. J. Ferroelectric polymers. *Science* **220**, 1115–1121 (1983).

10. Ramadan, K. S., Sameoto, D. & Evoy, S. A review of piezoelectric polymers as functional materials for electromechanical transducers. *Smart Mater. Struct.* **23**, 033001 (2014).
11. Kolda, R. R. & Lando, J. B. The effect of hydrogen-fluorine defects on the conformational energy of polytrifluoroethylene chains. *J. Macromol. Sci. B* **11**, 21–39 (1975).
12. Cais, R. E. & Kometani, J. M. Synthesis of pure head-to-tail poly(trifluoroethylenes) and their characterization by 470-MHz fluorine-19 NMR. *Macromolecules* **17**, 1932–1939 (1984).
13. Jaffe, B., Roth, R. S. & Marzullo, S. Piezoelectric properties of lead zirconate-lead titanate solid-solution ceramics. *J. Appl. Phys.* **25**, 809–810 (1954).
14. La-Orauttapong, D. et al. Phase diagram of the relaxor ferroelectric $(1-x)\text{Pb}(\text{Zn}_{1/3}\text{Nb}_{2/3})-\text{xPbTiO}_3$. *Phys. Rev. B* **65**, 144101 (2002).
15. Noheda, B., Cox, D. E., Shirane, G., Guo, J. & Ye, Z.-G. Phase diagram of the ferroelectric relaxor $(1-x)\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})-\text{xPbTiO}_3$. *Phys. Rev. B* **66**, 054104 (2002).
16. Farmer, B. L., Hopfinger, A. J. & Lando, J. B. Polymorphism of poly(vinylidene fluoride): potential energy calculations of the effects of head-to-head units on the chain conformation and packing of poly(vinylidene fluoride). *J. Appl. Phys.* **43**, 4293–4303 (1972).
17. Lovinger, A. J., Davis, D. D., Cais, R. E. & Kometani, J. M. The role of molecular defects on the structure and phase transitions of poly(vinylidene fluoride). *Polymer* **28**, 617–626 (1987).
18. Lovinger, A. J., Davis, G. T., Furukawa, T. & Broadhurst, M. G. Crystalline forms in a copolymer of vinylidene fluoride and trifluoroethylene (52/48 mol%). *Macromolecules* **15**, 323–328 (1982).
19. Davis, G. T., Furukawa, T., Lovinger, A. J. & Broadhurst, M. G. Structural and dielectric investigation on the nature of the transition in a copolymer of vinylidene fluoride and trifluoroethylene (52/48 mol %). *Macromolecules* **15**, 329–333 (1982).
20. Zhang, Q. M., Bharti, V. & Zhao, X. Giant electrostriction and relaxor ferroelectric behavior in electron-irradiated poly(vinylidene fluoride-trifluoroethylene) copolymer. *Science* **280**, 2101–2104 (1998).
21. Furukawa, T. & Takahashi, Y. Ferroelectric and antiferroelectric transitions in random copolymers of vinylidene fluoride and trifluoroethylene. *Ferroelectrics* **264**, 1739–1748 (2001).
22. Furukawa, T. & Seo, N. Electrostriction as the origin of piezoelectricity in ferroelectric polymers. *Jpn. J. Appl. Phys.* **29**, 675–680 (1990).
23. Katsouras, I. et al. The negative piezoelectric effect of the ferroelectric polymer poly(vinylidene fluoride). *Nat. Mater.* **15**, 78–84 (2016).
24. Broadhurst, M. G. & Davis, G. T. Physical basis for piezoelectricity in PVDF. *Ferroelectrics* **60**, 3–13 (1984).
25. Tashiro, K., Takano, K., Kobayashi, M., Chatani, Y. & Tadokoro, H. Structure and ferroelectric phase transition of vinylidene fluoride-trifluoroethylene copolymers: 2. VDF 55% copolymer. *Polymer* **25**, 195–208 (1984).
26. Bellet-Amalric, E. & Legrand, J. F. Crystalline structures and phase transition of the ferroelectric P(VDF-TrFE) copolymers, a neutron diffraction study. *Eur. Phys. J. B* **3**, 225–236 (1998).
27. Lovinger, A. J. & Cais, R. E. Structure and morphology of poly(trifluoroethylene). *Macromolecules* **17**, 1939–1945 (1984).
28. Bohlén, M. & Bolton, K. Conformational studies of poly(vinylidene fluoride), poly(trifluoroethylene) and poly(vinylidene fluoride-co-trifluoroethylene) using density functional theory. *Phys. Chem. Chem. Phys.* **16**, 12929–12939 (2014).
29. Wicker, A., Berge, B. & Lajzerowicz, J. Nonlinear optical investigation of the bulk ferroelectric polarization in a vinylidene fluoride/trifluoroethylene copolymer. *J. Appl. Phys.* **66**, 342–349 (1989).

Acknowledgements This research was funded by the US Office of Naval Research (grants N000141612082 and N000141612459). The supercomputer time at the National Center for Supercomputing Applications (NSF OCI-0725070 and ACI-1238993) was provided by NSF grant ACI-1615114. Y.L. thanks N. Wonderling, J. Stapleton and J. Long for technical assistance.

Reviewer information Nature thanks T. Kimura, C. Park and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.L. and Q.W. designed the research. W.X. was responsible for the polymer synthesis and NMR measurements. Y.L. prepared the polymer films and collected XRD and DMA data. H.A. performed FTIR measurements. Y.L. and H.A. carried out DSC and dielectric measurements. Y.L., H.A. and W.X. measured $P-E$ loops and electromechanical properties. B.Z., W.L. and J.B. performed the first-principle calculations and analysis. Y.L. and Q.W. wrote the manuscript and all authors revised the manuscript. Q.W. supervised the research.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0550-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0550-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Q.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Synthesis of P(VDF-TrFE) copolymers. De-ionized water and potassium peroxodisulfate initiator (from VWR) were added into a stainless steel vessel (PARR 452HC; 300 ml). The vessel was sealed and degassed via a vacuum pump and was subsequently cooled in a liquid nitrogen bath. Gaseous VDF and TrFE (SynQuest Laboratory Inc.) were pumped into the reaction vessel successively. The ratio between VDF and TrFE was controlled by tuning the pumping rates of VDF and TrFE. The vessel was then heated at 90 °C for 12 h. After the reaction was complete, the vessel was cooled to room temperature and opened. The precipitates in the vessel were washed by de-ionized water and methanol and then dried in vacuum. After several rounds of washing, the final white solid obtained was identified as P(VDF-TrFE). NMR spectra were acquired by a Bruker CDPX-300 spectrometer (300 MHz). Samples were dissolved in acetonitrile- d_3 (VWR) and scanned 128 times to reduce the noise in the spectra. The ^1H NMR spectra were used to determine the molar ratio of VDF and TrFE in the copolymers. The regiosequences and stereosequences in the ^{19}F NMR spectra were assigned according to the method described in previous works^{12,30} and are summarized and compared with previous results³¹ in Extended Data Fig. 2. Our synthesized polymers are nearly identical to commercial samples with the same composition in terms of chemical microstructure (regio- and stereo-sequences) and piezoelectric responses (Supplementary Figs. 4, 24, Supplementary Table 1).

Solution-cast P(VDF-TrFE) films. P(VDF-TrFE) was dissolved in *N,N*-dimethylformamide (DMF; Sigma-Aldrich) at a concentration of 120 mg ml⁻¹. The solution was stirred at a rate of 700 r.p.m. overnight and subsequently cast onto glass plates and dried at 70 °C for 16 h in a vacuum oven. Subsequently, the films were peeled off from the glass plates and annealed at 130 °C for 12 h. The typical film thickness was about 60 μm .

Structural and conformational characterization. XRD θ - 2θ scans at room temperature were performed using a PANalytical X'pert Pro MPD diffractometer in the Bragg-Brentano geometry with a source containing both Cu K α_1 and Cu K α_2 (with wavelengths of $\lambda_{\text{K}\alpha_1} = 1.54059 \text{ \AA}$ and $\lambda_{\text{K}\alpha_2} = 1.54442 \text{ \AA}$). Rietveld refinement was done using the Jade XRD analysis software (MDI). The atomic coordinates of P(VDF-TrFE) used in this work were built on the basis of previous studies²⁵. The structure was further optimized by taking into account the alternating deflections away from the planar zigzag chain structure to accommodate the steric hindrance between adjacent fluorine atoms³². The symmetry (monoclinic space group *Cc*) was imposed onto the whole structure ($C_{\text{VDF}} = 55 \text{ mol\%}$). The whole pattern fitting method included in the Jade software was used, which refines the lattice parameters, profile shape, scale factor and temperature factor. The agreement factor *R* is defined as

$$R = \sqrt{\frac{\sum [(I_o - I_c)^2 / I_o]}{\sum (I_o^2 / I_o)}}$$

where I_o is the experimental intensity and I_c is the calculated intensity. Another agreement factor, E_R , is defined as

$$E_R = \frac{N - Z}{\sum (I_o^2 / I_o)}$$

where *N* is the number of fitted data points and *Z* is the number of refined parameters. The ratio of R/E_R is called the 'goodness of fit'. An ideal refinement results in $R/E_R = 1$.

Temperature-dependent XRD data were obtained in the wide-angle X-ray scattering (WAXS) mode using a Xeuss 2.0 SAXS/WAXS system (50 kV, 0.60 mA). The system was in transmission geometry with a monochromated Cu K α_1 source (wavelength $\lambda = 1.54189 \text{ \AA}$) and a detector (Dectris Pilatus3R 200K). The polymer film was mounted using two thin Kapton plates (thickness of about 20 μm) and fixed in a Linkam HFX350 temperature cell in a Xeuss 2.0 sample stage. The role of the Kapton plate was to avoid any possible adherence to the cell when the temperature approached the melting temperature of the polymer samples. In this respect, Kapton acted as a background because its XRD characteristic peaks are located at much lower 2θ values ($2\theta \approx 5^\circ$). The Linkam HFX350 temperature-control stage was mounted vertically for transmission WAXS. The heating and cooling ramp rates were 1 °C min⁻¹. After the system reached the target temperature, the temperature was held constant for 10 min to maintain the equilibrium state. The WAXS measurement was then performed.

FTIR spectra were collected in the attenuated total reflectance mode using a Bruker Vertex V70 spectrometer equipped with a ZnSe crystal. The assignments of characteristic infrared absorbance bands were made on the basis of previous works^{19,33,34}. For temperature-dependent measurements, a round heating plate was embedded in a recycled-water system as a temperature-control stage to heat

and cool the attached polymer films. The recycled-water system was introduced into the setup to enhance temperature stability. The centre of the heating plate was attached to a K-type thermocouple (Omega) positioned very close to the samples. To control and monitor the measurement temperature, a heater controller (HARRICK; 200 °C) was used, which was connected to the heating plate and the thermocouple. All the spectra presented in Fig. 1f, g, 3d, e and Extended Data Figs. 3, 9 were obtained after reaching temperature stability in the whole setup.

DSC and DMA measurements. A DSC system (TA Q100) in the temperature-modulated mode was used to conduct thermal studies on the solution-cast films from 0 °C to 200 °C at a heating and cooling rate of 5 °C min⁻¹. The latent heat was obtained by integrating the exothermic peak related to the transition to the paraelectric phase, and the degree of crystallinity ΔX_c was calculated according to $\Delta X_c = \Delta H_m / [\omega(\text{VDF})\Delta H_0]$, where ΔH_m is the melting enthalpy of the polymer, calculated by integrating the melting peak in the DSC heating scan, $\omega(\text{VDF})$ is the mass ratio of VDF in the P(VDF-TrFE) copolymers and ΔH_0 is the melting enthalpy of a 100% crystalline poly(vinylidene fluoride) (PVDF) ($\Delta H_0 = 103.4 \text{ J g}^{-1}$)³⁵.

A DMA analyser (TA RSA-G2) in the tension mode was used to measure changes in the mechanical and viscoelastic properties of the polymer films as a function of temperature at 10 Hz with a heating and cooling rate of 2 °C min⁻¹. Young's modulus, *Y*, was determined from the slope of the initial part of a stress-strain curve measured by tensile testing at a constant linear rate of 0.01 mm s⁻¹.

Electrical and strain measurements. Gold electrodes of a typical thickness of 60 nm and a diameter of 4 mm were sputtered (Denton Vacuum, Desk IV) on both sides of the polymer films for the electrical measurements.

Dielectric spectra were acquired over a broad temperature range using a dielectric E4980A precision LCR meter (Keysight) in conjunction with a Delta Design oven (model 9023). The data were recorded at a heating rate of 1.5 °C min⁻¹ from 25 °C to 100 °C in the frequency range 100 Hz–1 MHz. In this temperature range, the dielectric relaxation measured in the copolymers was mainly attributed to molecular motions in the crystalline regions and was independent of the amorphous regions³⁶.

For the *P-E* measurements, a modified Sawyer-Tower circuit was used, in which the electroded copolymer films were subjected to a triangular bipolar wave. Electric-field-induced strain data were collected simultaneously by a special test fixture with a resolution better than 3 nm and a frequency range of 0.01–10 Hz. After the *P-E* measurements, poled polymer films were obtained. To measure the strain response simultaneously with the *P-E* curves, a homemade strain fixture with a lock-in amplifier (Stanford Research 830) was designed. The electroded sample was held in place sandwiched by two probes, similarly to a typical d_{33} meter. The bottom probe was the electrical high and had a miniature high-voltage connector. The baseplate of the strain fixture was connected to the ground on the voltage supply or equipment rack using the baseplate ground wire. The strain fixture consisted of a small frame, a linear variable differential transformer and the hold-down micrometer, and the whole fixture with the connecting wires was immersed in Fluorinert to prevent air breakdown. The strain fixture allowed us to measure a small motion from the sample with high sensitivity, with the output signal quantifying the electric-field-induced displacement from the polymer films.

Computational modelling. The structural energies of various P(VDF-TrFE) copolymers were calculated with the DFT software Quantum ESPRESSO³⁷ and ultrasoft pseudopotentials from the Standard Solid State Pseudopotentials (SSSP) library. Nonlocal intermolecular van der Waals interactions were included through the VDW-DF2 functional³⁸. The wavefunction cutoff energy was 55 rydbergs (Ry; 1 Ry = 13.605 eV). The Monkhorst-Pack *k*-point mesh 2-2-2 was used³⁹. The Brodyen-Fletcher-Goldfarb-Shanno method was employed to calculate structural relaxations.

The copolymer chains with a dominant VDF-TrFE segment were modelled as (VDF-TrFE)_{*m*}-(VDF)_{*n*} or (VDF-TrFE)_{*m*}-(TrFE)_{*n*}, whereas those with a dominant TrFE-TrFE segment were modelled as (TrFE)_{*m*}-(VDF)_{*n*}. The structures were initialized with their torsional angles being 180° (T), 60° (G) or -60° (\bar{G}). Two chain conformations were considered, the all-*trans* planar and the 3/1-helical ((TG)₃) conformations. We modelled the isotactic and syndiotactic configurations by arranging the fluorine atoms in -CHF- groups along the chains either on the same side or alternately. For each VDF content and for the two sets of models (with dominant (VDF-TrFE) or (TrFE-TrFE) segments), the energy difference between the all-*trans* and 3/1-helical structures was calculated using the energy of the syndiotactic all-*trans* structures as the reference.

The main goal of the DFT calculations was to determine the energy differences between the 3/1-helix and all-*trans* structures with $C_{\text{VDF}} = 40$ –60 mol%. In our DFT approach, the calculations were performed for copolymers with different VDF concentrations in a supercell. In this respect, the total energies of copolymers with different compositions varied considerably in our calculations. Therefore, we compared their relative—rather than absolute—energies (Fig. 4) and chose the total

energy of the syndiotactic planar-*trans* structure as the reference so that the energy differences between different phases are highlighted clearly.

Theoretical models of piezoelectricity in P(VDF-TrFE) copolymers. The physical description of the negative longitudinal piezoelectric effect in the P(VDF-TrFE) copolymers requires the understanding of the origin of negative piezoelectricity, which is still in dispute despite extensive research for nearly 50 years. Generally, three different models have been proposed. The first one is the so-called dimensional model, which explains the piezoelectric effect in terms of the deformation of the amorphous regions^{24,40}, and in which the dipoles are assumed to be fixed and d_{33} can be estimated through $d_{33} \approx -P_r/Y$, where P_r is the remanent polarization. The second one is the electrostriction model⁴¹, which typically describes d_{33} in normal ferroelectrics according to $d_{33} = 2Q_{33}\epsilon_r\epsilon_0P_r$, where ϵ_r and ϵ_0 are the relative and vacuum permittivities, respectively, and Q_{33} is the electrostrictive coefficient. Electrostriction refers to dimension changes in response to an applied electric field, which are due to the internal stress caused by the force of the electric field on charges^{41,42}. In contrast to the piezoelectric effect (which has a linear dependence on the field), electrostriction scales with E^2 and does not depend on the field direction, and the strain S_3 can be determined^{22,23} according to $S_3 = Q_{33}P^2$. Recalling that electrostriction occurs in both crystalline and amorphous regions⁴², we note that the formula $d_{33} = 2Q_{33}\epsilon_r\epsilon_0P_r$ generally works for normal ferroelectrics⁴¹, in which ferroelectricity arises only from the crystalline regions. This analysis disagrees with previous work²², which attributed electrostriction to the dimensional effect arising from the amorphous regions of the P(VDF-TrFE) copolymers. The electrostrictive explanation summarized here is also supported by recent first-principles calculations on PVDF⁴³. As a consequence, the dimensional and electrostrictive models actually contradict each other. The third model calculates the electromechanical contribution from crystalline–amorphous interfacial regions on the basis of the electrostriction ($2Q_{33}\epsilon_r\epsilon_0P_r$) model in the crystalline regions²³.

In summary, the contribution of the crystalline domains of the P(VDF-TrFE) copolymers to the negative piezoelectricity depends on the ratio between the measured d_{33} and the calculated $2Q_{33}\epsilon_r\epsilon_0P_r$ ^{22,23}. For instance, the electrostrictive model is dominant when $d_{33}/(2Q_{33}\epsilon_r\epsilon_0P_r) = 1$. Otherwise, additional contributions—that is, the crystalline–amorphous interfacial coupling ($|d_{33}/(2Q_{33}\epsilon_r\epsilon_0P_r)| > 1$)—have to be considered²³.

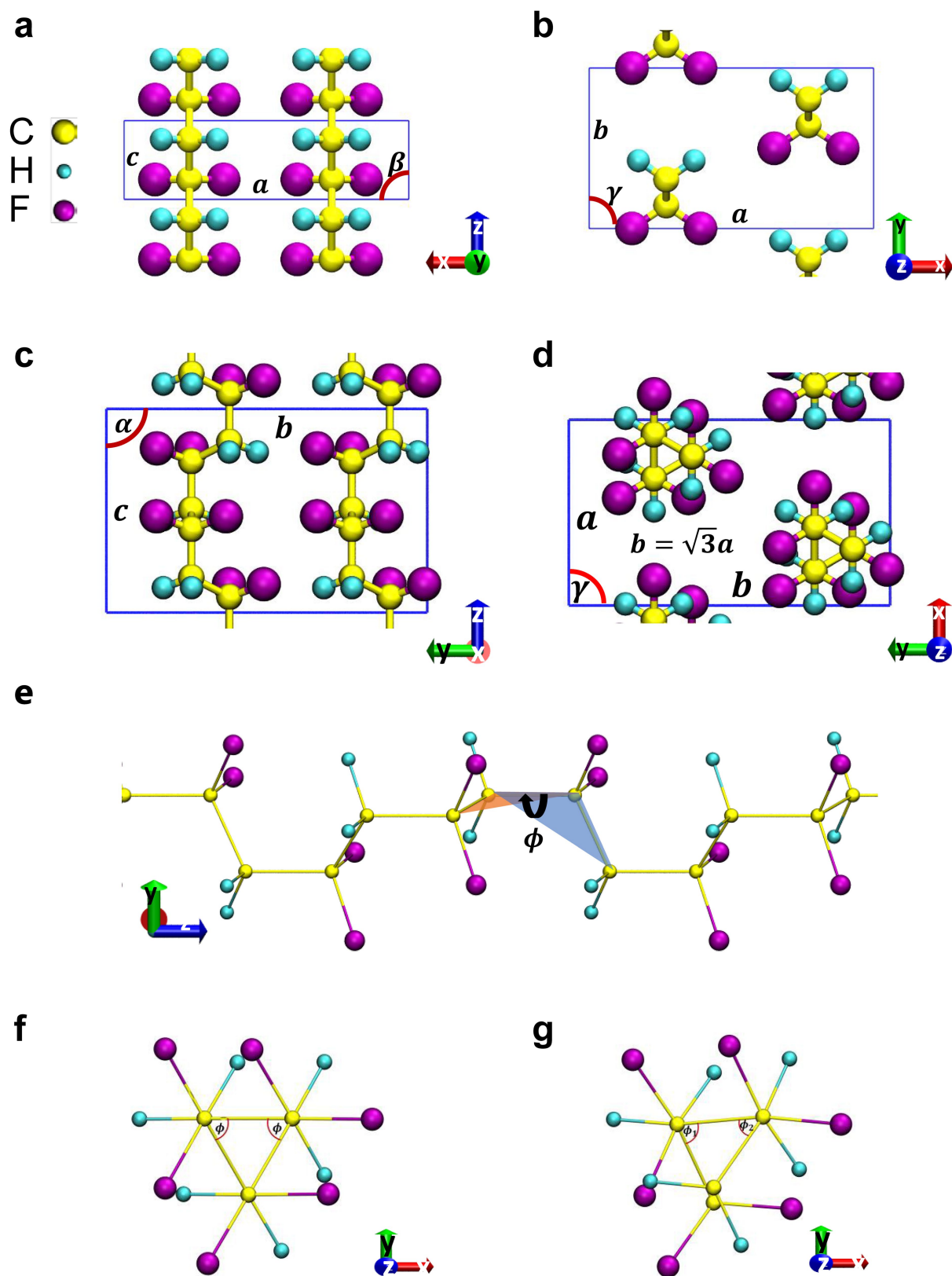
Especially important here is the electrostrictive coefficient Q_{33} because all the other parameters, such as ϵ_r , ϵ_0 and P_r , can be measured directly. However, almost all of the previous works^{22,23} extracted Q_{33} at room temperature, which may involve extrinsic contributions from ferroelectric switching and domain-wall motion in the ferroelectric phase. To use electrostrictive coefficients obtained at room temperature, one needs further high-temperature measurements in the paraelectric phase. However, such a route is technically challenging for ferroelectric polymers that are limited by a lossy paraelectric phase, especially at a lower frequency (1 Hz) and a reduced breakdown field. The true Q_{33} cannot be obtained in a lossy paraelectric phase because of the extrinsic contributions from the dramatic increase of the electrical conductivity, even though an S_3 – P response is obtained. In our case, we chose a typical morphotropic composition with $x_{\text{VDF}} = 50$ mol% for P(VDF-TrFE)

and measured simultaneously the strain and the polarization response at 70 °C, which is just above the Curie temperature of about 65 °C. For high-temperature strain and polarization measurements, the strain fixture was set inside a Delta Design oven (model 2300) containing a moderate amount of Fluorinert to fully immerse the fixture. A K-type thermocouple (Omega) was buried in the Fluorinert, positioned close to the samples, to monitor the measurement temperature. High-temperature data were collected only after the temperature had equilibrated over the whole setup.

Data availability

The data supporting the findings of this study are available within the paper and its Supplementary Information.

- Soulesin, T., Ladmira, V., Lannuzel, T., Domingues Dos Santos, F. & Améduri, B. Importance of microstructure control for designing new electroactive terpolymers based on vinylidene fluoride and trifluoroethylene. *Macromolecules* **48**, 7861–7871 (2015).
- Yagi, T. & Tatemoto, M. A fluorine-19 NMR study of the microstructure of vinylidene fluoride-trifluoroethylene copolymers. *Polym. J.* **11**, 429–436 (1979).
- Hasegawa, R., Takahashi, Y., Chatani, Y. & Tadokoro, H. Crystal structures of three crystalline forms of poly(vinylidene fluoride). *Polym. J.* **3**, 600–610 (1972).
- Kobayashi, M., Tashiro, K. & Tadokoro, H. Molecular vibrations of three crystal forms of poly(vinylidene fluoride). *Macromolecules* **8**, 158–171 (1975).
- Kim, K. J., Reynolds, N. M. & Hsu, S. L. Spectroscopic analysis of the crystalline and amorphous phases in a vinylidene fluoride/trifluoroethylene copolymer. *Macromolecules* **22**, 4395–4401 (1989).
- Gomes, J., Serrado Nunes, J., Sencadas, V. & Lanceros-Mendez, S. Influence of the β -phase content and degree of crystallinity on the piezo- and ferroelectric properties of poly(vinylidene fluoride). *Smart Mater. Struct.* **19**, 065010 (2010).
- Furukawa, T., Ohuchi, M., Chiba, A. & Datela, M. Dielectric relaxations and molecular motions in homopolymers and copolymers of vinylidene fluoride and trifluoroethylene. *Macromolecules* **17**, 1384–1390 (1984).
- Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
- Lee, K., Murray, É. D., Kong, L., Lundqvist, B. I. & Langreth, D. C. Higher-accuracy van der Waals density functional. *Phys. Rev. B* **82**, 081101 (2010).
- Pack, J. D. & Monkhorst, H. J. 'Special points for Brillouin-zone integrations'—a reply. *Phys. Rev. B* **16**, 1748–1749 (1977).
- Broadhurst, M. G., Davis, G. T., McKinney, J. E. & Collins, R. E. Piezoelectricity and pyroelectricity in polyvinylidene fluoride—a model. *J. Appl. Phys.* **49**, 4992–4997 (1978).
- Lines, M. E. & Glass, A. M. *Principles and Applications of Ferroelectrics and Related Materials* (Oxford Univ. Press, Oxford, 1977).
- Jaffe, B., Cook, W. R. & Jaffe, H. *Piezoelectric Ceramics* (Academic Press, London, 1971).
- Bystrov, V. S. et al. Molecular modeling of the piezoelectric effect in the ferroelectric polymer poly(vinylidene fluoride) (PVDF). *J. Mol. Model.* **19**, 3591–3602 (2013).



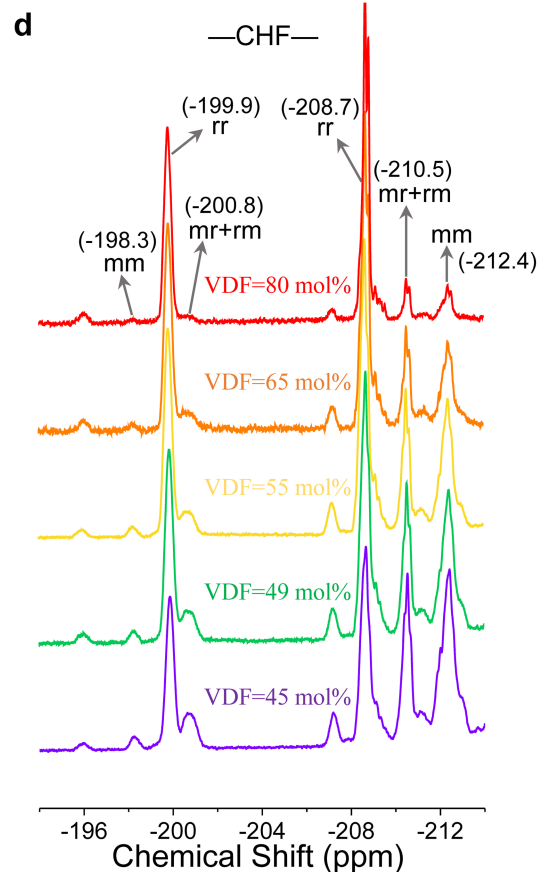
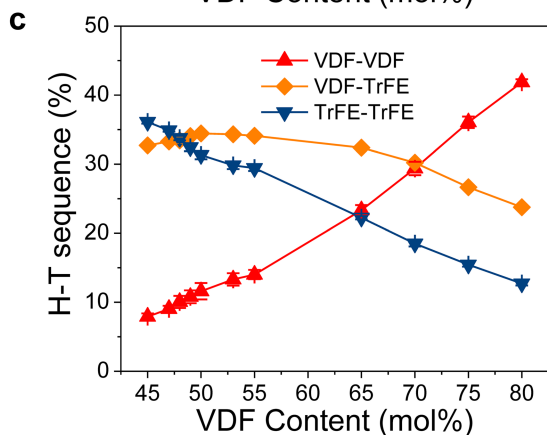
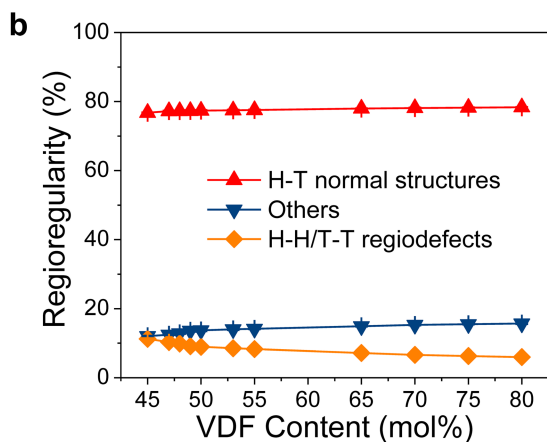
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Crystallographic structures of different phases in the phase diagram of P(VDF-TrFE) copolymers. **a, b**, Side (**a**) and top (**b**) view of the *trans*-planar phase. **c, d**, Side (**c**) and top (**d**) view of the 3/1-helical phase. **e**, Side view of the *gauche* dihedral angle ϕ in the 3/1-helical phase. **f**, Top view of the regular 3/1-helical phase ($\phi = 60^\circ$). **g**, Top view of the (pseudo-)disordered 3/1-helical phase ($\phi_1, \phi_2 \neq 60^\circ$). Intramolecular disorder occurs in the (pseudo-)disordered 3/1-helical phase in terms of random departure from the *gauche* dihedral angle of 60° in the regular form (**f**). The 3/1-helical phase becomes energetically more stable than the *trans*-planar phase as the VDF content decreases in the vicinity of 50 mol% (Fig. 4a). Moreover, intramolecular disorder develops mainly in the 3/1-helical phase at the critical VDF content of 55 mol%, while the fraction of the regular *trans*-planar phase decreases. With further decrease of the VDF content ($49 \text{ mol}\% \leq C_{\text{VDF}} \leq 55 \text{ mol}\%$), the fraction of the disordered 3/1-helical phase increases substantially

and strongly competes with the *trans*-planar phase (we call this phase coexistence of the pseudo-disordered and ordered phases, or a morphotropic phase in the phase diagram). Here the morphotropic phase ($49 \text{ mol}\% \leq C_{\text{VDF}} \leq 55 \text{ mol}\%$) differs from the disordered paraelectric phase owing to the existence of a phase transition (Fig. 3). As the VDF content decreases below 49 mol%, only the disordered 3/1-helical phase (namely, the pseudo-disordered phase) is energetically preferred. In this case, the pseudo-disordered phase has nearly the same structure as the disordered paraelectric phase owing to the absence of a phase transition (Fig. 3), and differs from both the morphotropic and the VDF-rich compositions. We note that we use the structure of PVDF to simplify this discussion. For the case of P(VDF-TrFE), the substitution of some H atoms by some F atoms leads to different chain tacticities of P(VDF-TrFE) (Fig. 1a), which requires a slight generalization of the above description.

a

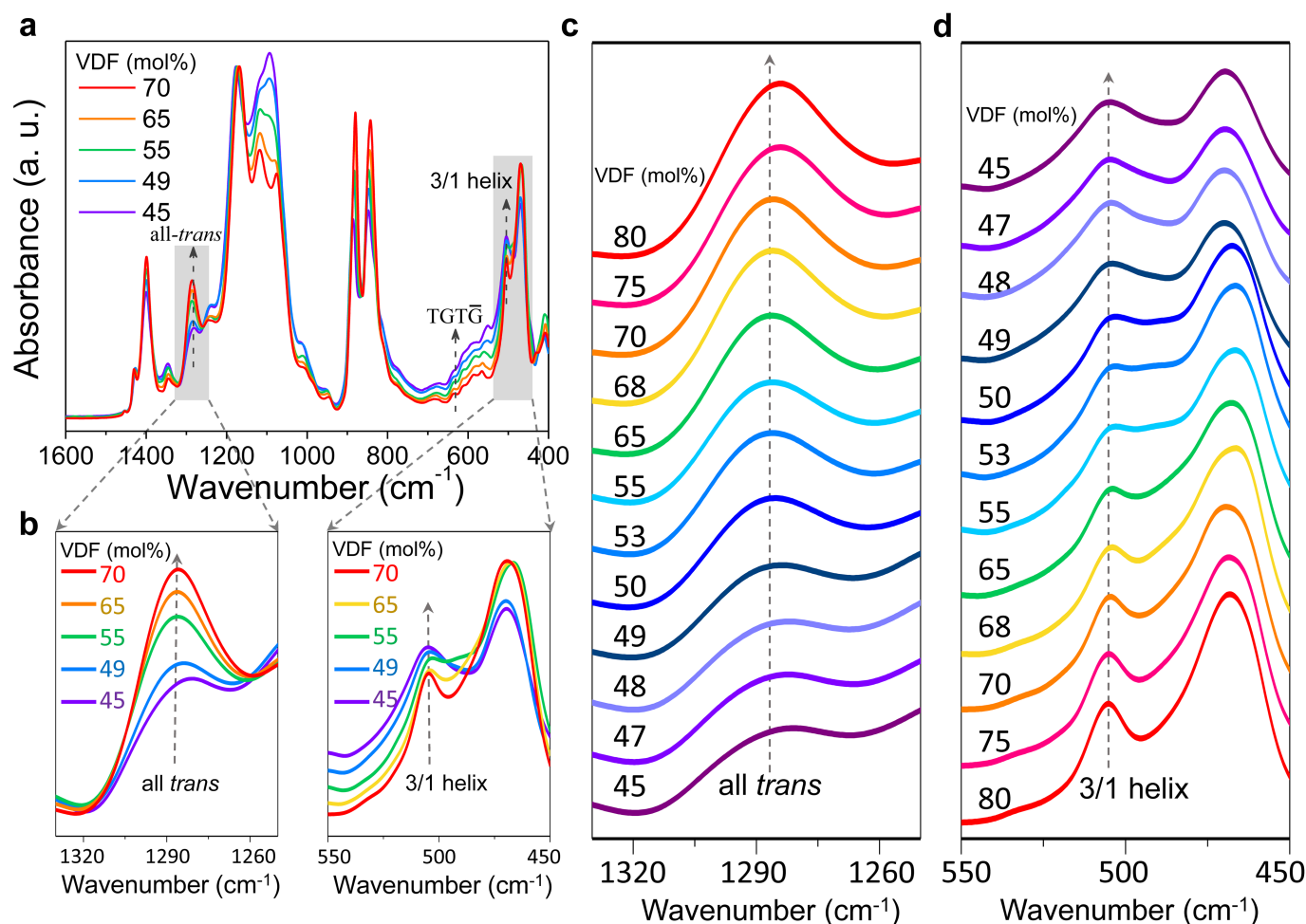
Regioregularity	5C Sequence	Designation	Chemical Shift (ppm)
H-T or T-H	CF ₂ CH ₂ CF₂ CH ₂ CF ₂	VDF-VDF, H-T	-93.2
	CF ₂ CH ₂ CF₂ CHFCF ₂	VDF-TrFE, H-T	-107.8
	CF ₂ CHFC CF₂ CHFCF ₂	TrFE-TrFE, H-T	-119.5 to -124.8
	CHFC CF₂ CHFCF ₂ CHF	TrFE-TrFE, T-H	-207.3 to -213.5
	CH ₂ CF ₂ CHFCF₂ CH ₂	TrFE-VDF, T-H	-197.5 to -201.5
H-H/T-T or T-T/H-H	CH ₂ CH ₂ CF₂ CF ₂ CH ₂	VDF-VDF-VDF, T-T/H-H	-117.2
	CHFC CHFCF₂ CF ₂ CHF	TrFE-TrFE-TrFE, T-T/H-H	-124.8 to -130.0
	CH ₂ CHFC CF₂ CF ₂ CH ₂	VDF-TrFE-VDF, T-T/H-H	-131.1
	CF ₂ CF ₂ CHFCF₂ CHFCF ₂	VDF-TrFE-TrFE, H-H/T-T	-218.5 to -220.4
Others	CHFC CH₂CF₂ CH ₂ CF ₂	TrFE-VDF-VDF, T-T/ H-T	-94.8 to -95.8
	CH ₂ CH ₂ CF₂ CH ₂ CF ₂	VDF-VDF-VDF, T-T/ H-T	-96.3 to -97.8
	CF ₂ CH ₂ CF₂ CF ₂ CHF	VDF-VDF-TrFE, H-T/H-H	-114.2



Extended Data Fig. 2 | See next page for caption.

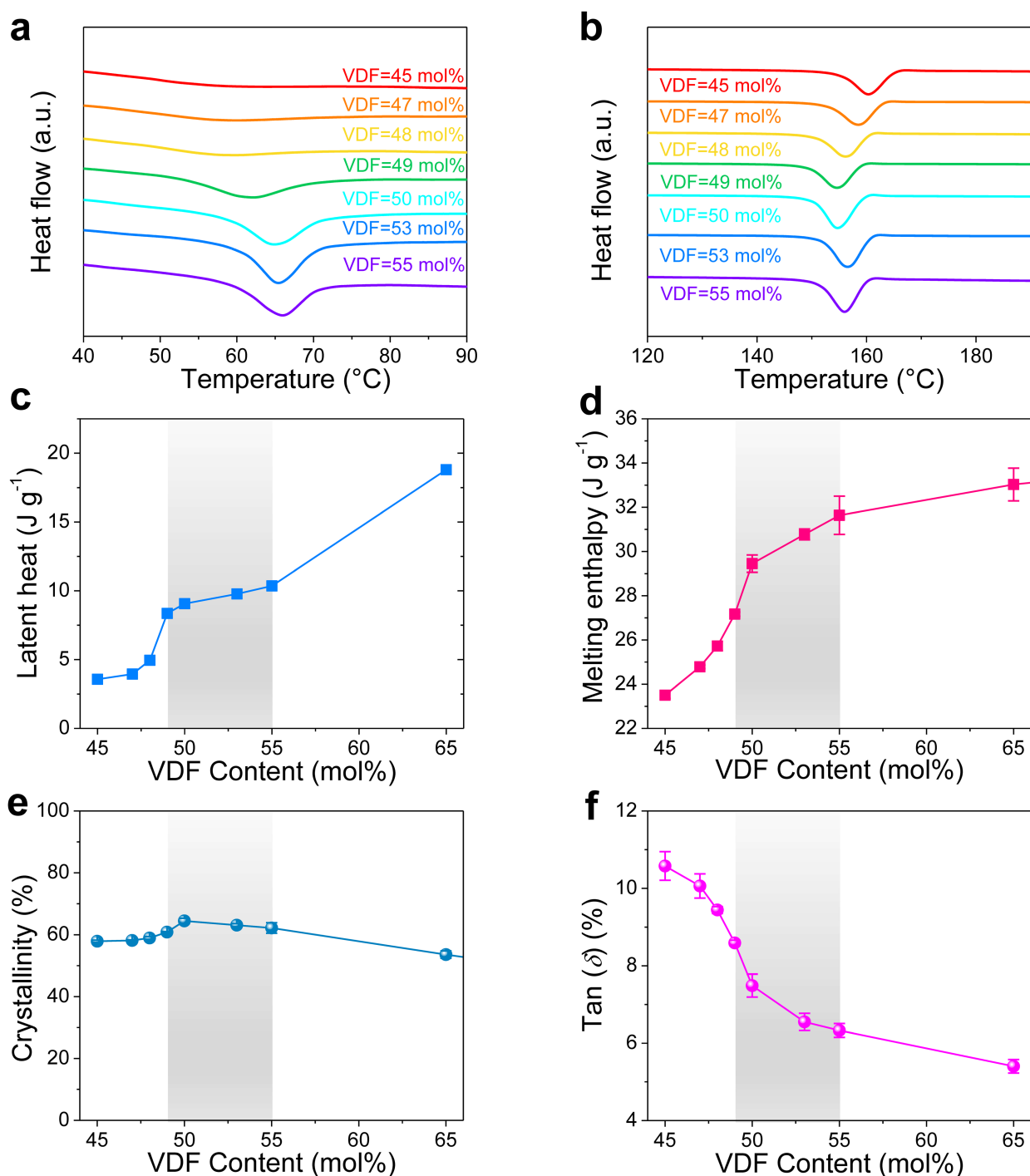
Extended Data Fig. 2 | Analysis of ^{19}F NMR spectra of P(VDF-TrFE) copolymers. **a**, Assignments of ^{19}F NMR signals of P(VDF-TrFE). The monomers indicated by the ^{19}F NMR signals are undelined. H-H, head to head; H-T, head to tail; T-T, tail to tail. **b**, Unconditional probabilities of different regiosequences as a function of VDF content. P(VDF-TrFE) is dominated by normal H-T isoregic sequences (about 76.7%–78.3%) with almost constant H-H/T-T regiodefects and non-regioisomers (marked as ‘others’). The slight variation in regioirregular sequences indicates that neither regiodefects nor non-regioisomers are responsible for the observed conformational competition. Instead, we identify the dominant contribution to be the dramatic tacticity change upon the formation of the MPB-like transition. **c**, Unconditional probabilities of normal H-T sequences consisting of the VDF-VDF, VDF-TrFE and TrFE-TrFE segments. With decreasing VDF content, the TrFE-TrFE units grow substantially, becoming even larger than their VDF-TrFE

counterparts (which remain nearly constant for $C_{\text{VDF}} = 45\text{--}65\text{ mol\%}$) for $C_{\text{VDF}} < 49\text{ mol\%}$, and this change is accompanied by a remarkable decrease in the VDF-VDF units. These results clearly indicate that the polymer chain becomes more PTrFE-like for the copolymers with TrFE-rich compositions. Our findings disagree with previous results³¹, which showed the H-T VDF-TrFE sequence to be predominant for the copolymers (for $C_{\text{VDF}} \approx 30\text{--}75\text{ mol\%}$). The poor NMR resolution (56.5 MHz) of the previous measurement led to the absence of many resonance peaks (above 130 p.p.m.). Moreover, only the $-\text{CF}_2-$ resonance area was considered, and the contributions from the $-\text{CHF}-$ resonance area were disregarded³¹. Consequently, the regioregularity was not appropriately described and the chain tacticity was not analysed. Error bars in **b** and **c** represent standard deviations obtained from at least three measurements using different samples and are typically smaller than the symbols. **d**, The $-\text{CFH}-$ resonances. All the triad assignments are indicated by grey arrows.



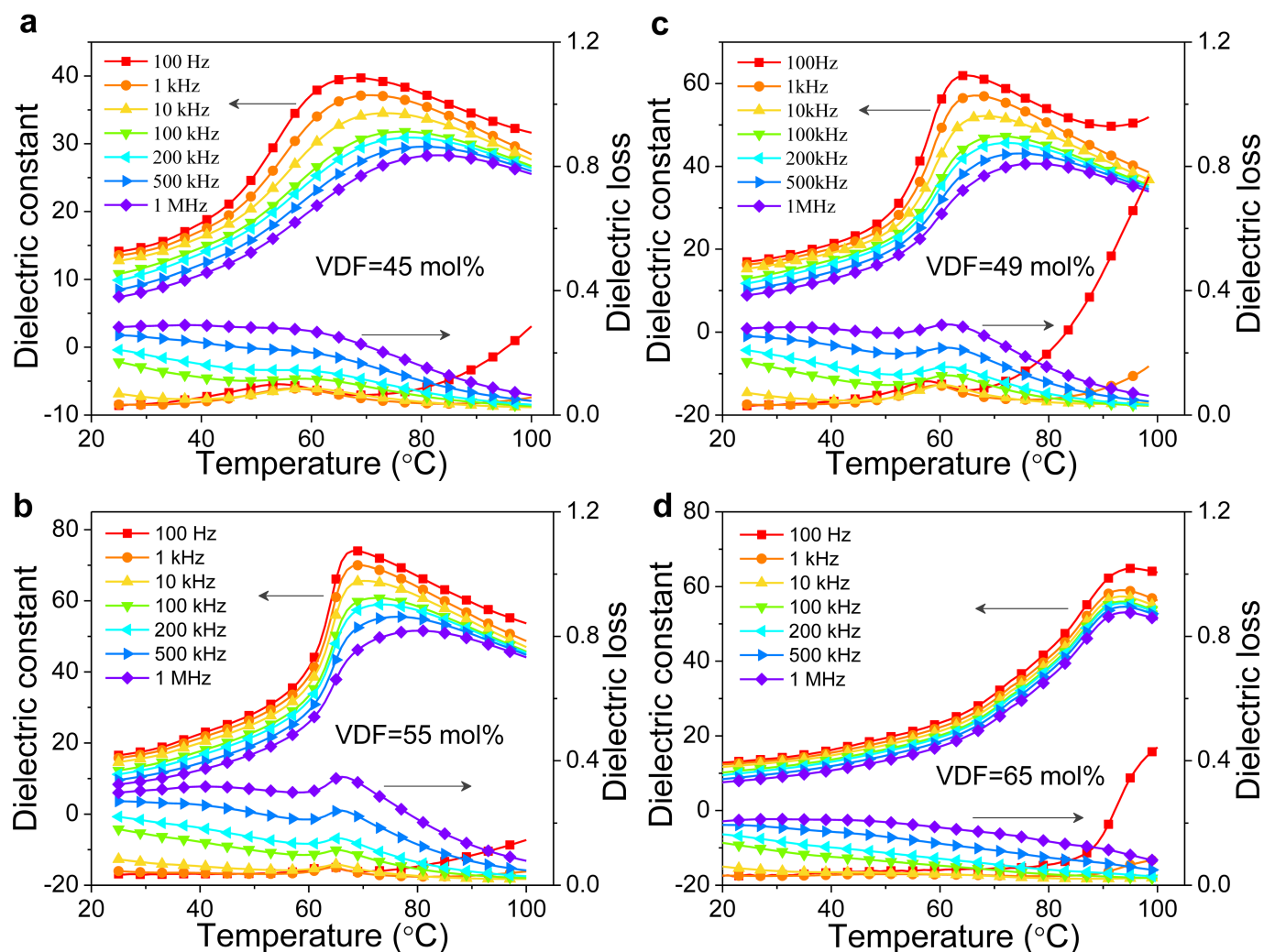
Extended Data Fig. 3 | FTIR spectra of P(VDF-TrFE) copolymers at room temperature. **a**, Raw data for selected polymer compositions ($C_{\text{VDF}} = 70$ mol%, 65 mol%, 55 mol%, 49 mol% and 45 mol%). The intensities of the characteristic bands corresponding to the TGTG conformation near 614 cm^{-1} and to the 3/1-helical conformation at around 507 cm^{-1} increase considerably with decreasing VDF content. Because of the extremely low band intensity of the TGTG conformation,

we mainly consider the conformational interconversion between the all-*trans* and 3/1-helical conformations. **b**, Zoom-in of the grey regions in **a**. The left panel shows the band characteristic of the all-*trans* conformation at around $1,290\text{ cm}^{-1}$ and the right panel presents the band characteristic of the 3/1-helical conformation at around 507 cm^{-1} . **c**, **d**, FTIR spectra of the magnified regions around $1,290\text{ cm}^{-1}$ and 507 cm^{-1} , with appropriate offset adjustments for clarity.



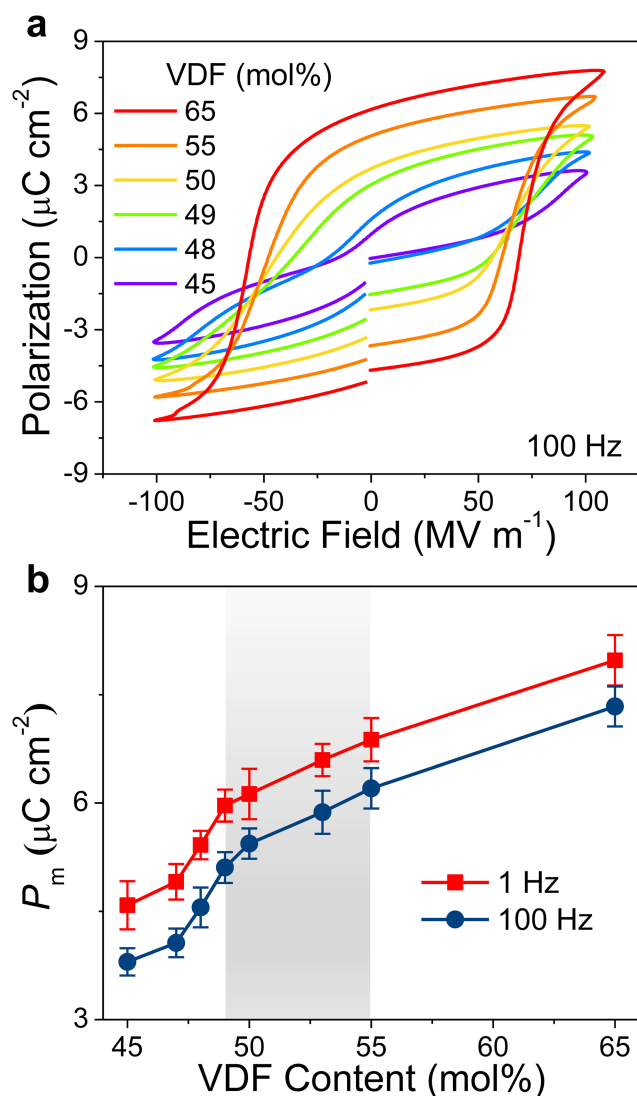
Extended Data Fig. 4 | DSC and DMA results for P(VDF-TrFE) copolymers. **a, b**, Heat flow versus temperature. **c**, Latent heat. **d**, Melting enthalpy. **e**, Crystallinity. **f**, $\tan(\delta)$ peak. $\tan(\delta)$ is defined as the ratio of the loss modulus to the storage modulus, measured using oscillatory shear stress tests, and can be used to evaluate the ratio between the viscous and elastic components per cycle of sample deformation. A very broad exothermic peak is found for $C_{\text{VDF}} < 49$ mol% (**a**), which is accompanied by a dramatic reduction in latent heat (**c**). As C_{VDF} decreases from 55 mol% to 49 mol%, the relative change in crystallinity (**e**) and the shift of the transition temperature are limited to about 2% and 6%, respectively, which cannot explain the observed approximately 60% enhancement in d_{33} achieved at $C_{\text{VDF}} = 50$ mol%. We find that $\tan(\delta)$ decreases considerably

with increasing VDF content (**e**), which implies that P(VDF-TrFE) copolymers with high VDF content are more elastic. This might be due to lattice contraction with increasing VDF content (Fig. 1e), which imposes restrictions against molecular mobility. We measured the elastic modulus, known as Young's modulus Y , to provide a measure of stiffness for our polymers (Extended Data Fig. 7e). The Y value of the P(VDF-TrFE) copolymers increases substantially with increasing VDF content, mediated by an intermediate (grey) region ($49 \text{ mol}\% \leq C_{\text{VDF}} \leq 55 \text{ mol}\%$). The compositional evolutions of the latent heat and $\tan(\delta)$ reveal the existence of an intermediate transition region (grey). Error bars in **c–f** represent standard deviations obtained from at least three measurements using different samples.

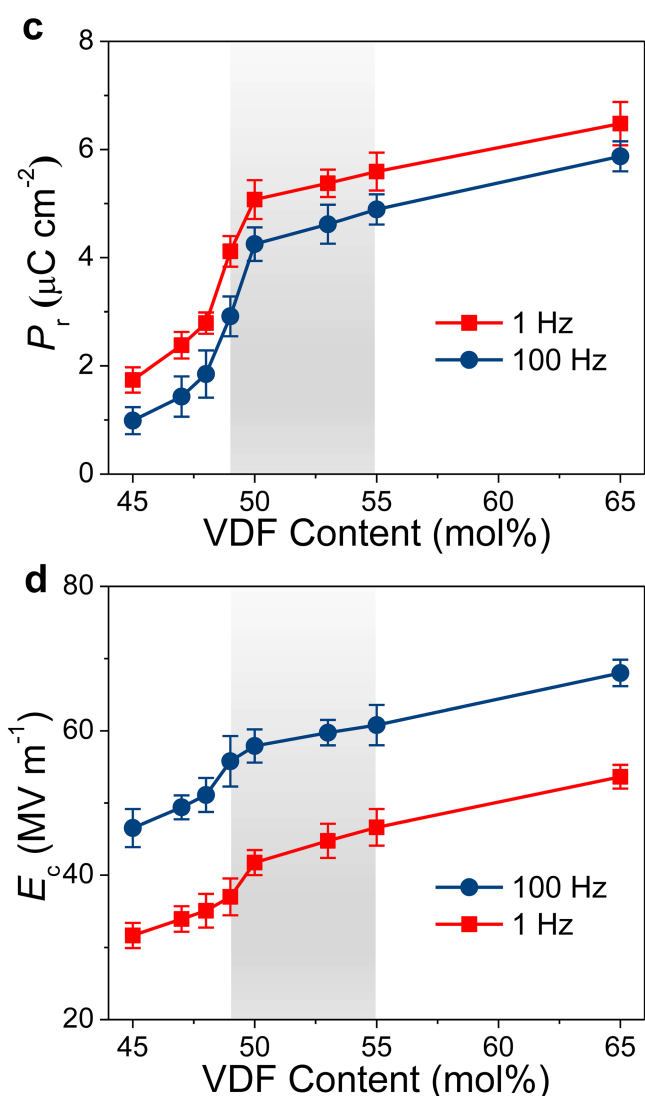


Extended Data Fig. 5 | Temperature-dependent dielectric constant and loss of P(VDF-TrFE) copolymers. **a–d**, Dielectric constant as a function of temperature for P(VDF-TrFE) copolymers with $C_{\text{VDF}} = 45$ mol% (**a**) $C_{\text{VDF}} = 49$ mol% (**b**), $C_{\text{VDF}} = 55$ mol% (**c**) and $C_{\text{VDF}} = 65$ mol% (**d**). **a–c** show relaxor behaviours, whereas **d** shows the typical dielectric response of normal ferroelectrics, in which the dielectric peak is independent of frequency. The dielectric response near T_{max} results from

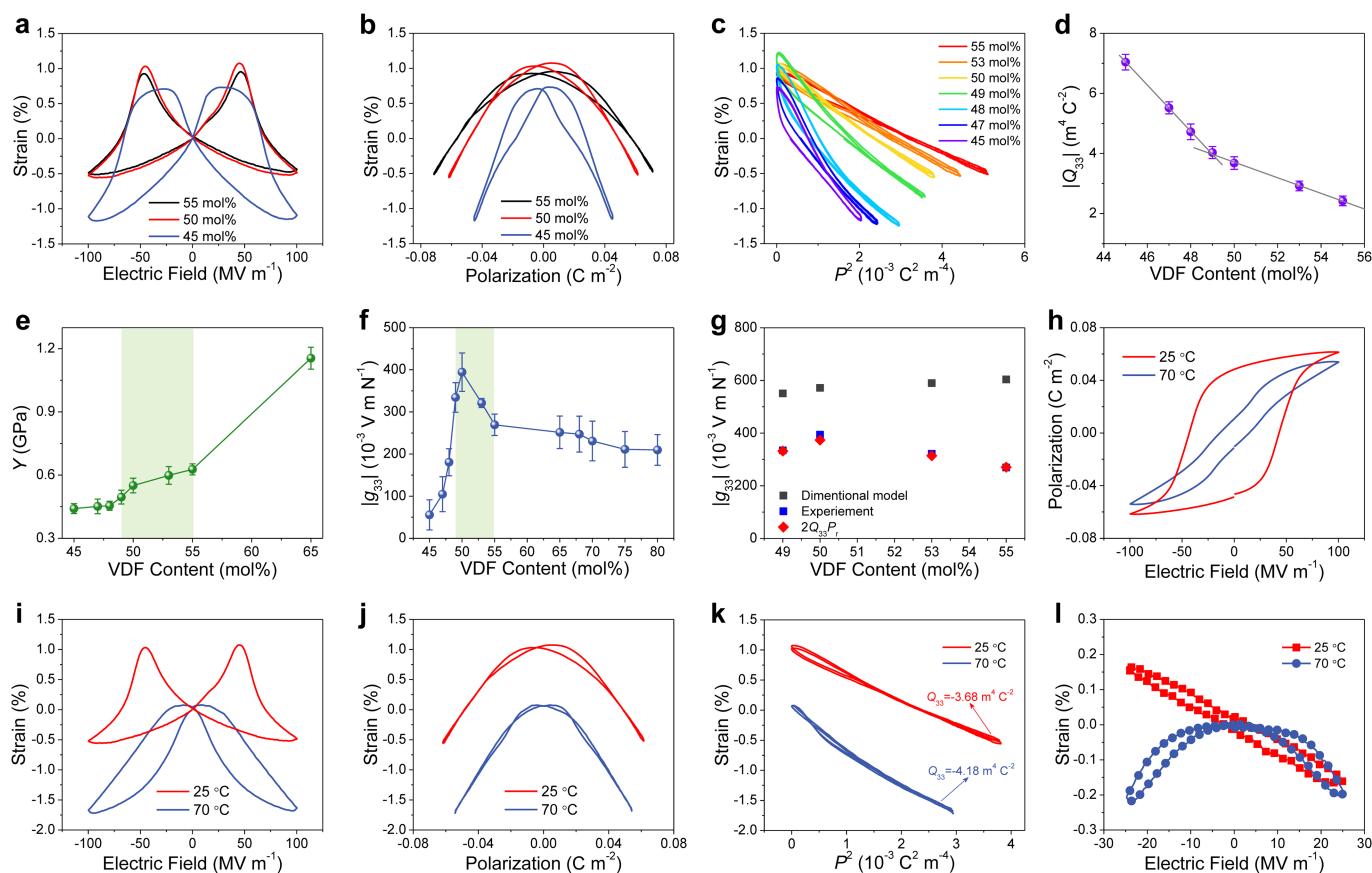
molecular motions in the crystalline regions rather than in the amorphous regions³². Therefore, both normal-ferroelectric and relaxor properties coexist in the crystalline regions of P(VDF-TrFE) within the transition region ($49 \text{ mol}\% \leq C_{\text{VDF}} \leq 55 \text{ mol}\%$). The discovery of relaxor behaviour not only explains the emergence of the peak at low 2θ in the XRD pattern (Fig. 1d) but also allows the correlation of its origin with intramolecular disorder. The arrows in **a–d** are guides for the eyes.



Extended Data Fig. 6 | Polarization hysteresis in P(VDF-TrFE) copolymers at room temperature. **a**, Polarization versus electric field, measured using a triangular a.c. electric field of 100 Hz. A polarization hysteresis curve—characteristic of a normal ferroelectric—is shown for $C_{\text{VDF}} \geq 49$ mol%, whereas an antiferroelectric-like hysteresis loop is observed for $C_{\text{VDF}} < 49$ mol%. **b–d**, Maximum polarization P_m (**b**),



remanent polarization P_r (**c**) and coercive field E_c (**d**) as a function of VDF content, measured at 1 Hz and 100 Hz. Considerable reduction in P_r is found in **c** ($C_{\text{VDF}} < 49$ mol%), which is indicative of the disappearance of long-range ferroelectric order. Error bars in **b–d** represent standard deviations obtained from at least three measurements using different samples.

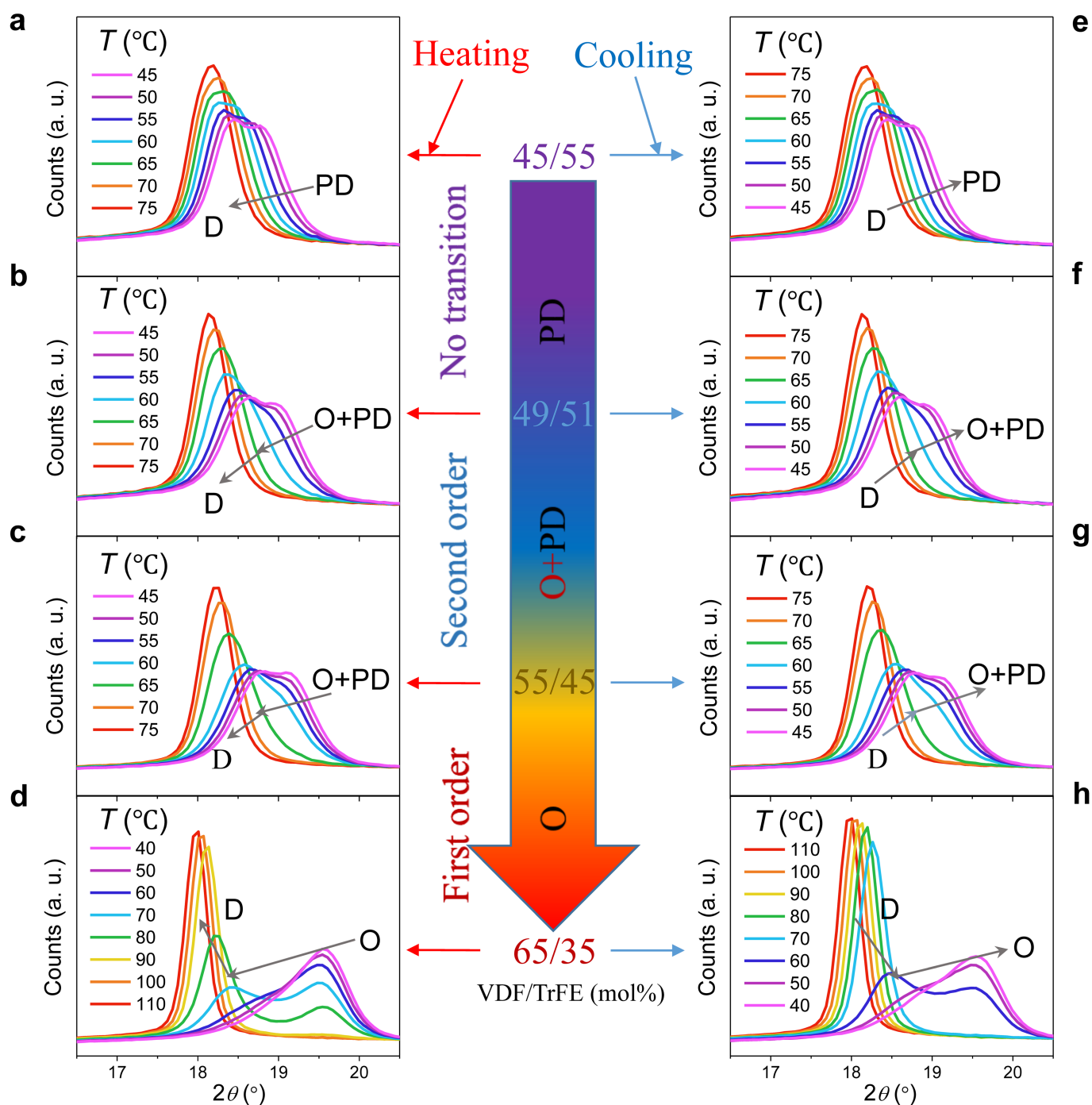


Extended Data Fig. 7 | Electrostrictive properties and theoretical descriptions of the piezoelectric effect in P(VDF-TrFE) copolymers.

a, Strain–electric field (S_3 – E) curves. **b**, Strain–polarization (S_3 – P) curves. **c**, S_3 – P^2 . Typical P(VDF-TrFE) compositions are shown in **a**–**c** for clarity. **d**, Q_{33} as a function of VDF content. **e**, Young's modulus Y as a function of VDF content. **f**, The magnitude of the piezoelectric voltage constant, g_{33} , as a function of VDF content at room temperature. g_{33} is determined according to $g_{33} = d_{33}/(\epsilon_r \epsilon_0)$, where ϵ_r and ϵ_0 are the relative and vacuum permittivities, respectively. The lines in **d**–**f** are guides for the eyes. The light-green shaded areas in **e** and **f** indicate the transition

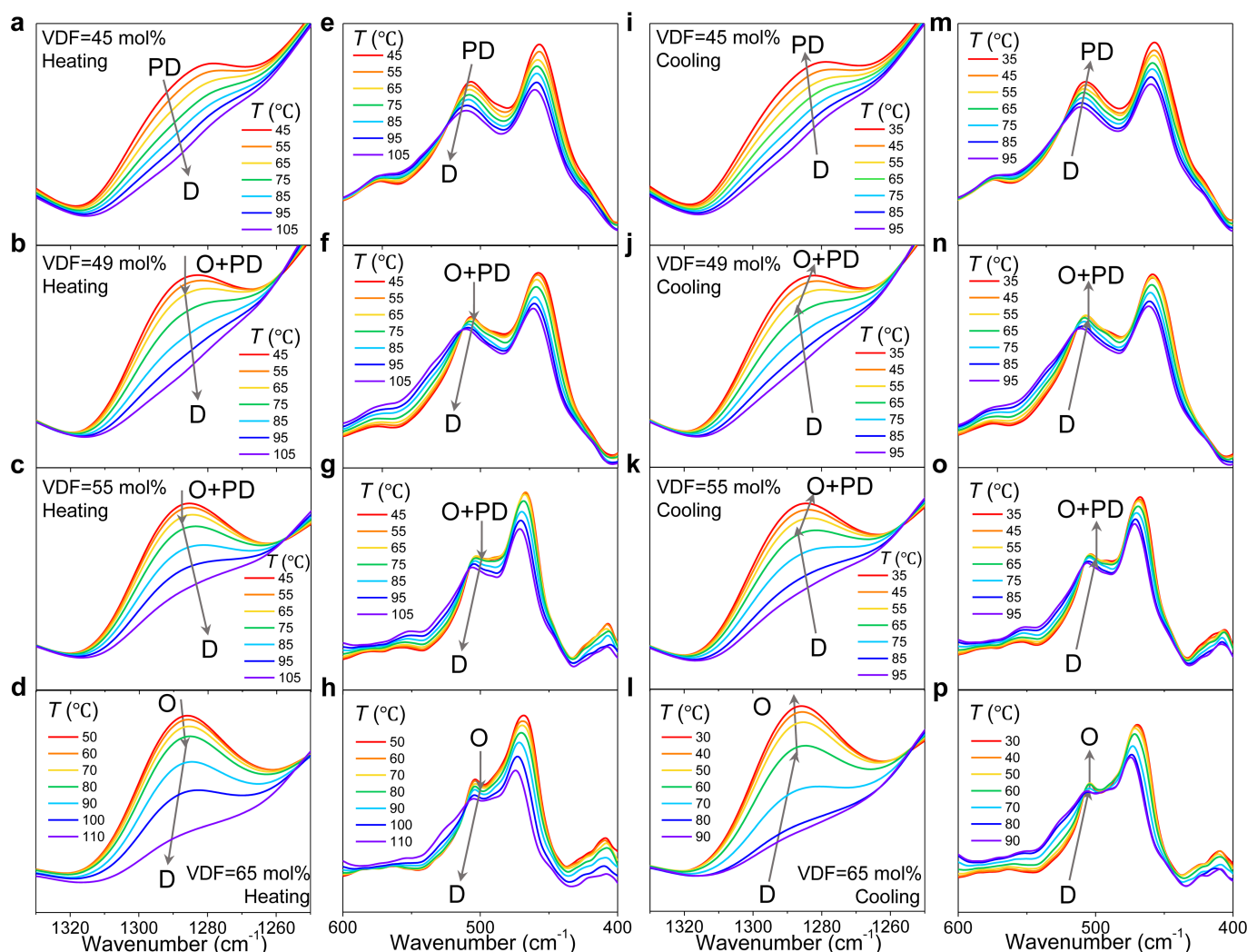
region. **g**, Comparison between experimental and theoretical g_{33} values.

h–**i**, Electrostrictive data for P(VDF-TrFE) copolymer at 50/50 mol% measured at 70 °C. **h**, P – E loops. **i**, Electric-field-induced strain. **j**, S_3 – P curves. **k**, S_3 – P^2 curves. **l**, Electric-field-induced strain at an electric field of 25 MV m^{−1}. The calculated d_{33} is -68.3 pC N^{−1} using $Q_{33} = -4.18$ m⁴ C^{−2} (**k**), which is slightly larger in magnitude than the experimental data (-63.5 ± 3.2 pC N^{−1}) for P(VDF-TrFE) with $C_{\text{VDF}} = 50$ mol%. A detailed discussion about the data shown here can be found in Supplementary Information. Error bars in **d**–**f** represent standard deviations obtained from at least ten measurements using different samples.



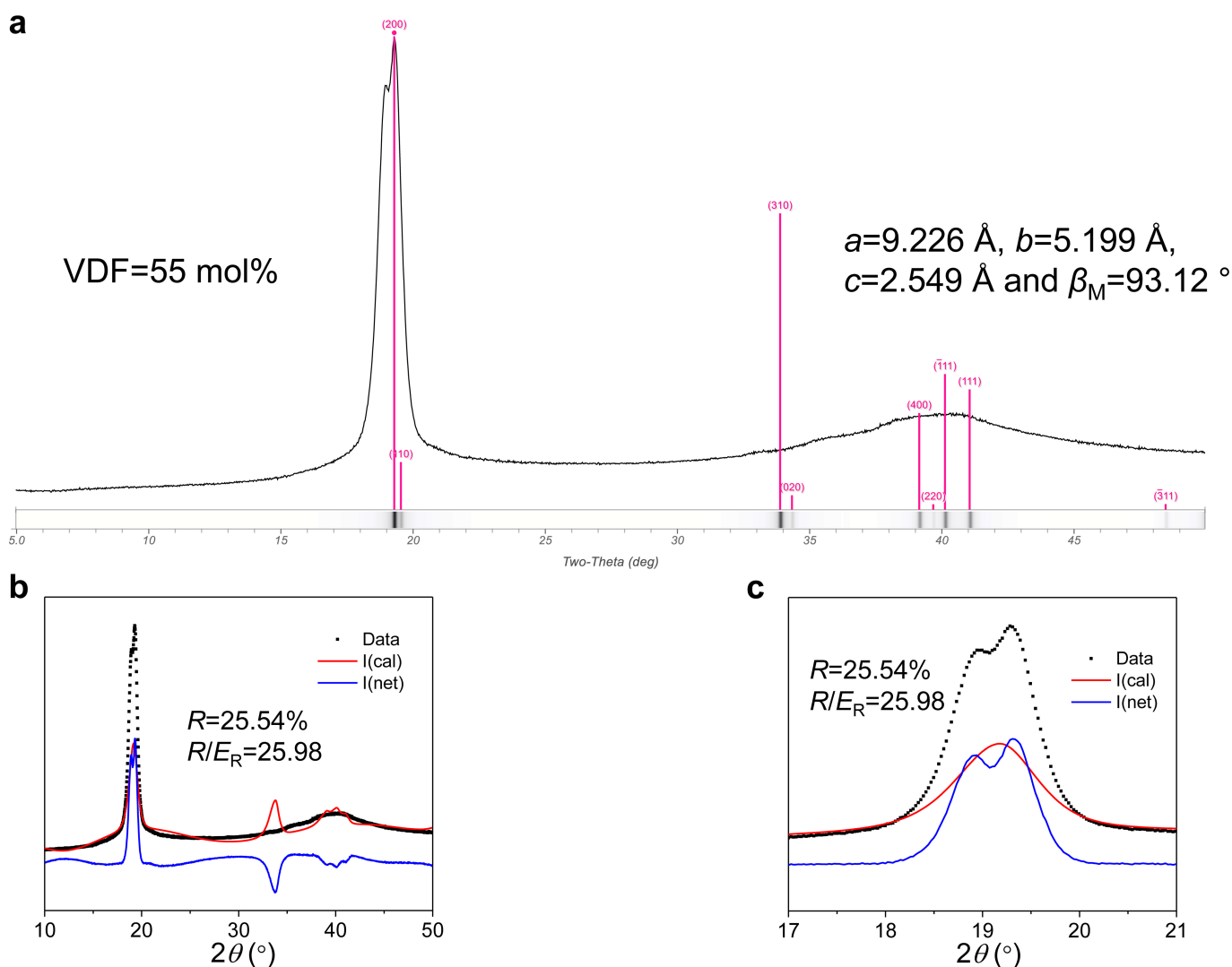
Extended Data Fig. 8 | X-ray patterns of P(VDF-TrFE) copolymers at various temperatures. **a–h**, X-ray patterns of P(VDF-TrFE) copolymers with selected VDF contents (**a** and **e**, $C_{\text{VDF}} = 45$ mol%; **b** and **f**, $C_{\text{VDF}} = 49$ mol%; **c** and **g**, $C_{\text{VDF}} = 55$ mol%; **d** and **h**, $C_{\text{VDF}} = 65$ mol%) upon heating (**a–d**) and cooling (**e–h**). 'O', 'D' and 'PD' indicate the ordered, disordered and pseudo-disordered phases (Extended Data Fig. 1), respectively. For $C_{\text{VDF}} < 49$ mol%, the peaks shift towards a lower 2θ at

an even rate upon heating (**a**) and cooling (**e**), indicating the absence of a phase transition. For $C_{\text{VDF}} \geq 49$ mol%, a change in shift rates is observed near the phase transition temperature (**b–d**, **f–h**), indicating the emergence of the order–disorder phase transition. Moreover, the order of the phase transition can be determined by analysing the thermal hysteresis at the transition temperature from the heating and cooling processes.



Extended Data Fig. 9 | Infrared absorbance bands of P(VDF-TrFE) copolymers at around $1,290\text{ cm}^{-1}$ and 507 cm^{-1} , measured at various temperatures. **a–h**, Infrared absorbance curves obtained during heating for copolymers with $C_{\text{VDF}} = 45\text{ mol\%}$ (**a**, **e**), $C_{\text{VDF}} = 49\text{ mol\%}$ (**b**, **f**), $C_{\text{VDF}} = 55\text{ mol\%}$ (**c**, **g**) and $C_{\text{VDF}} = 65\text{ mol\%}$ (**d**, **h**). **i–p**, Corresponding curves obtained during cooling for $C_{\text{VDF}} = 45\text{ mol\%}$ (**i**, **m**), $C_{\text{VDF}} = 49\text{ mol\%}$ (**j**, **n**), $C_{\text{VDF}} = 55\text{ mol\%}$ (**k**, **o**) and $C_{\text{VDF}} = 65\text{ mol\%}$ (**l**, **p**). ‘O’, ‘D’ and ‘PD’ represent the ordered, disordered and pseudo-disordered phases (Extended Data Fig. 1), respectively. For $C_{\text{VDF}} < 49\text{ mol\%}$, the

characteristic bands that correspond to both the all-*trans* and 3/1-helical conformations shrink or expand uniformly upon heating (**a**, **e**) or cooling (**i**, **m**), implying no phase transition. For $C_{\text{VDF}} \geq 49\text{ mol\%}$, two distinct regimes can be clearly resolved (as marked by grey arrows), signifying the presence of the order–disorder phase transition. Thermal hysteresis is negligible for $49\text{ mol\%} \leq C_{\text{VDF}} \leq 55\text{ mol\%}$ (second-order phase transition), whereas substantial thermal hysteresis is observed for P(VDF-TrFE) with a VDF content of 65 mol\% (first-order phase transition).



Extended Data Fig. 10 | Rietveld analysis for P(VDF-TrFE) copolymer with $C_{VDF}=55 \text{ mol\%}$. **a**, XRD pattern in the 2θ range $[10^\circ, 50^\circ]$, measured at room temperature. The indexing was done according to the monoclinic space group Cc . a , b and c are the refined lattice parameters and β_M is the monoclinic angle. **b**, **c**, Fitting results. Apparent disagreements are observed for the (200,110) and (310) peaks (**a**), leading to a large agreement factor of $R=25.54\%$ (**b**, **c**), which is within the range (about 23%–29%) reported in a previous study²⁵. This poor fitting result is only slightly improved (about 23.03%; Supplementary Fig. 8) by considering chain deflection³². These results suggest that simply considering the planar-zigzag-based chain structure (namely, the ‘cooled’ phase model²⁵) cannot yield a reasonable fit for the XRD data near the transition region. Additionally, no reasonable fit was obtained by

considering other space groups, such as Cm or $Cm2m$ (not shown here). Indeed, a large agreement factor (about 10%–30%) was usually achieved in previous structural refinements^{25,26,32} in PVDF and its copolymers. One of the main reasons for this disagreement is that polymer crystallography generally corresponds to the ideal condition, disregarding concomitant structural defects (disorder) and complex morphology (a composite of amorphous, crystalline and intermediate regions) that exist in real polymers. Typically, the Bragg peaks are very broad at high values of diffraction angle 2θ and the number of available reflections is limited for morphotropic compositions, which considerably increases the difficulty of crystallographic analysis. Obviously, further studies—both theoretical and experimental—of the crystal structure of P(VDF-TrFE) near the transition region are required.

Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages

Cheng Wang^{1,2,6}, Mian Zhang^{1,6}, Xi Chen³, Maxime Bertrand^{1,4}, Amirhassan Shams-Ansari^{1,5}, Sethumadhavan Chandrasekhar³, Peter Winzer³ & Marko Lončar^{1*}

Electro-optic modulators translate high-speed electronic signals into the optical domain and are critical components in modern telecommunication networks^{1,2} and microwave-photonic systems^{3,4}. They are also expected to be building blocks for emerging applications such as quantum photonics^{5,6} and non-reciprocal optics^{7,8}. All of these applications require chip-scale electro-optic modulators that operate at voltages compatible with complementary metal-oxide-semiconductor (CMOS) technology, have ultra-high electro-optic bandwidths and feature very low optical losses. Integrated modulator platforms based on materials such as silicon, indium phosphide or polymers have not yet been able to meet these requirements simultaneously because of the intrinsic limitations of the materials used. On the other hand, lithium niobate electro-optic modulators, the workhorse of the optoelectronic industry for decades⁹, have been challenging to integrate on-chip because of difficulties in microstructuring lithium niobate. The current generation of lithium niobate modulators are bulky, expensive, limited in bandwidth and require high drive voltages, and thus are unable to reach the full potential of the material. Here we overcome these limitations and demonstrate monolithically integrated lithium niobate electro-optic modulators that feature a CMOS-compatible drive voltage, support data rates up to 210 gigabits per second and show an on-chip optical loss of less than 0.5 decibels. We achieve this by engineering the microwave and photonic circuits to achieve high electro-optical efficiencies, ultra-low optical losses and group-velocity matching simultaneously. Our scalable modulator devices could provide cost-effective, low-power and ultra-high-speed solutions for next-generation optical communication networks and microwave photonic systems. Furthermore, our approach could lead to large-scale ultra-low-loss photonic circuits that are reconfigurable on a picosecond timescale, enabling a wide range of quantum and classical applications^{5,10,11} including feed-forward photonic quantum computation.

Future photonic systems require modulators with a CMOS-compatible drive voltage, a large bandwidth, a low optical insertion loss, a high extinction ratio, excellent signal quality and compatibility with large-scale manufacturing. Because discrete lithium niobate (LN) modulators are difficult to integrate, many other photonic platforms compatible with microfabrication processes have been pursued instead, including those based on silicon^{1,12,13}, indium phosphide^{14,15}, polymers^{16,17} and plasmonics¹⁸. These have shown excellent scalability and distinct performance merits, including the potential for integration with CMOS electronics (Si), low drive voltages (InP, polymer), ultra-high bandwidths (polymer, plasmonics) and small footprints (Si, plasmonics). Although the integration problem has been greatly alleviated in these platforms, a modulator that simultaneously meets all desired performance aspects remains elusive because of the non-ideal electro-optic properties of the underlying materials.

The material properties of LN are well suited for realizing ultra-fast modulation, low-voltage operation and low optical losses at the same

time. The strong electro-optic (Pockels) effect in LN leads to a linear change of its refractive index in response to an applied voltage, on femtosecond timescales¹⁹. Although it has been known for some time that microstructured LN devices can provide better modulator performance²⁰, most commercial LN modulators are still based on titanium-indiffusion or proton-exchange waveguides, because LN is notoriously difficult to etch⁹. These waveguides typically have a low refractive index contrast Δn of around 0.02 between core and cladding, resulting in a large optical mode size²¹. The weak optical confinement requires metal electrodes to be spaced far apart from the optical waveguide (about 10 μm), lowering the electro-optic efficiency. As a result, LN modulators today are much larger in size and require much higher drive voltages than the material is capable of supporting.

In recent years, the LN-on-insulator platform has emerged as a promising candidate for integrated high-performance modulators. In this approach, a single-crystal, submicrometre-thick LN film is bonded on top of a low-index substrate (silicon dioxide, SiO_2), and waveguides are created by dry etching the LN device layer²². This has led to a range of LN photonic devices with high index contrast of >0.7 and tightly confined optical modes^{23–29}. Electro-optic modulators with promising electro-optic efficiencies have been demonstrated^{25–27,29}. However, the actual switching voltages, bandwidths and optical losses in these demonstrations still suffer from critical trade-offs, limited by non-ideal etching, reduced overlap between electrical and optical fields, and/or the inefficient microwave signal delivery. Whether it is possible to simultaneously achieve a low on/off switching voltage, an ultra-high bandwidth and a low optical loss in LN modulators has remained an outstanding question.

Here we demonstrate monolithically integrated LN electro-optic modulators (Fig. 1) that overcome such trade-offs, featuring a switching voltage of 1.4 V while supporting very high bandwidths. Our integrated modulators operate in a travelling-wave Mach–Zehnder interferometer (MZI) configuration that uses highly confined co-propagating microwave and optical fields with matched group velocities and low propagation losses. A 50:50 Y-junction splits the input light into two LN optical waveguides that form the arms of the MZI. The optical waveguides run through the dielectric gaps of a ground-signal-ground coplanar microwave strip line (Fig. 1d). As a result, the microwave electric field has opposite signs across the two LN waveguides, thus inducing (through the Pockels effect) an optical phase delay on one arm and an optical phase advance on the other. This optical phase difference results in constructive/destructive interference at the output 50:50 Y-junction, and thereby an amplitude modulation of the output optical signal (Fig. 1c). An important figure of merit for MZI modulators is the half-wave voltage (V_π), defined as the voltage required to induce a π -phase difference between the two modulator arms, changing the optical transmission from maximum to minimum. For a device with 20-mm-long microwave strip line electrodes, we measure a low V_π of 1.4 V (Fig. 1c), which allows the modulator to be directly driven by a CMOS circuit. Importantly, our devices also

¹John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ²Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, China. ³Nokia Bell Labs, Holmdel, NJ, USA. ⁴LP2N, Institut d'Optique Graduate School, CNRS, University of Bordeaux, Talence, France. ⁵Department of Electrical Engineering and Computer Science, Howard University, Washington, DC, USA. ⁶These authors contributed equally: Cheng Wang, Mian Zhang. *e-mail: loncar@seas.harvard.edu

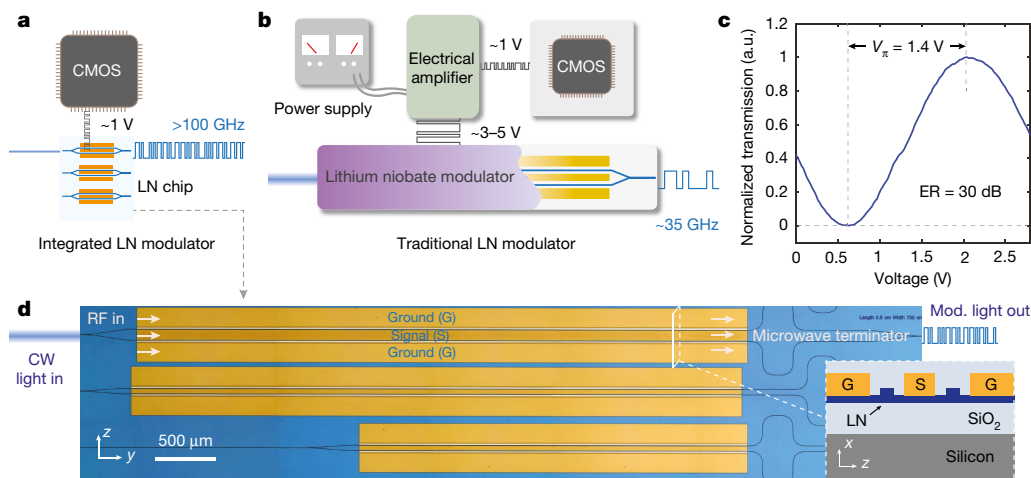


Fig. 1 | Nanophotonic LN modulators compatible with CMOS drive voltages. **a, b**, Schematic comparison of the data-transmitting set-ups for integrated (**a**) and traditional (**b**) LN modulators. The nanophotonic LN modulator (**a**) supports direct CMOS driving with high bandwidths (>100 GHz), while traditional modulators (**b**) require large and power-consuming electrical driver amplifiers and have limited bandwidths (approximately 35 GHz). **c**, Normalized optical transmission of a 20-mm

device as a function of the applied voltage, showing a low half-wave voltage of 1.4 V. The measured extinction ratio is 30 dB. **d**, Microscope image of the fabricated chip consisting of three Mach-Zehnder modulators with various microwave signal line widths and device lengths. The thin-film configuration allows for maximum field overlap and velocity matching between microwave and optics. Inset shows the cross-sectional schematic of the nanophotonic LN modulator. Mod., modulated; RF, radiofrequency.

feature a high optical power extinction ratio of about 30 dB between on and off states (see Methods).

In the travelling-wave electrode configuration, longer microwave strip lines could be used to induce a larger optical phase shift, thus reducing the V_π value. However, this degrades the electro-optic bandwidth, owing to exacerbated mismatch between microwave and optical velocities, and larger microwave loss. This contradictory requirement on electrode length results in a voltage–bandwidth trade-off⁹. In traditional LN modulators, the large optical mode size requires the metal electrodes to be placed far from the optical waveguides, and very long electrodes are required to reduce the voltage to even modest levels. As a result, the voltage–bandwidth performance of these modulators is typically limited to V_π about 3.5 V and bandwidth about 35 GHz, requiring power-consuming electrical amplifiers to drive them (Fig. 1b)⁹.

In contrast, the thin-film LN modulator can overcome the bandwidth–voltage performance limitation by maximizing the electro-optic overlap using photonic structures with a strong optical confinement^{25,26,29}. Owing to the increased modulation efficiency, our devices are much shorter (ranging from 5 mm to 20 mm) than conventional counterparts (typically >5 cm), while allowing for V_π values at CMOS levels (Fig. 1a). Moreover, as the optical mode is highly localized in the submicrometre waveguide region whereas the microwave mode resides largely in the substrate, it is possible to engineer the microwave and

optical group velocities independently, by designing the layer thicknesses of the LN/SiO₂/Si stack (see Methods). As a result, our devices can maintain velocity matching between the optical and the microwave signals at very high microwave frequencies without sacrificing the electro-optic overlap. Figure 2a presents the measured small-signal electro-optic response of the 20-mm-long device, showing a high 3-dB bandwidth of >45 GHz. These devices also possess ultra-low on-chip optical losses of <0.5 dB (see Methods)²⁸.

We use the low-voltage, high-bandwidth and low-loss integrated modulators to demonstrate data modulation at 70 Gbit s^{-1} directly driven by a CMOS circuit (Fig. 2b–d). High-speed electrical signals are generated by a CMOS digital-to-analogue conversion (DAC) circuit and directly used to drive our modulator without an electrical amplifier. Figure 2c, d shows measured constellation diagrams (Fig. 2c, d, left) from a coherent receiver, which recovers both amplitude and phase of the output optical field at each data-sampling instant. The vertical and horizontal axes correspond to the in-phase and quadrature components of the measured optical fields. Distinct constellations with fewer overlapping data points between ‘0’ and ‘1’ correspond to the desired lower bit-error ratios (BER). The eye diagrams (Fig. 2c, d right) are generated by up-sampling the received digital data for better visualization (see Methods). Low BERs correspond to clear separation between the discrete data levels at the sampling time (i.e. middle of the eye diagrams). At a peak-to-peak drive voltage (V_{pp}) of 200 mV, the modulated optical

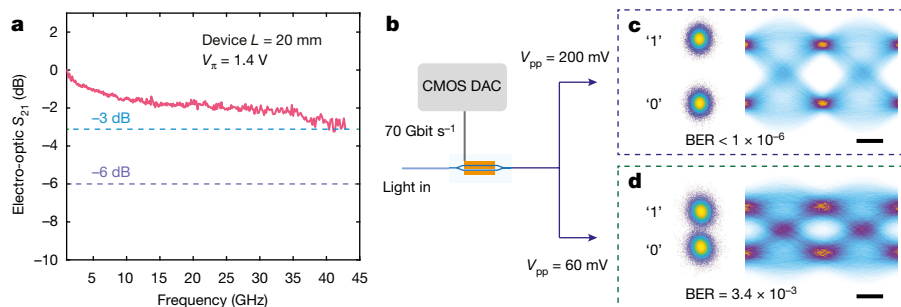


Fig. 2 | Directly CMOS-driven data transmission at 70 Gbit s^{-1} . **a**, Small-signal electro-optic response of a device with an active modulation length of 20 mm, showing a high 3-dB bandwidth of >45 GHz. S_{21} , transmission coefficient of the scattering matrix. **b**, The device is used for data-transmission experiments at a rate of 70 Gbit s^{-1} rate, directly

driven by a CMOS circuit. **c, d**, Measured constellation diagrams obtained with a coherent receiver (left), and the reconstructed eye diagrams (right). At peak-to-peak voltages of 200 mV (**c**) and 60 mV (**d**), the measured BERs are $<1 \times 10^{-6}$ and 3.4×10^{-3} , respectively. The eye diagrams are obtained by up-sampling the received data for better visualization. Scale bars, 5 ps.

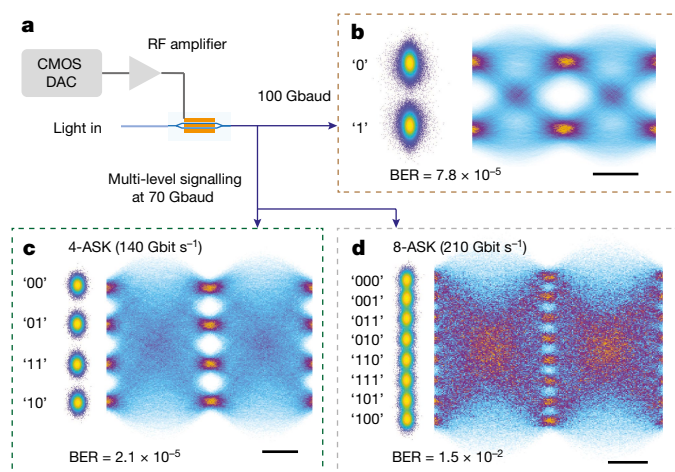


Fig. 3 | Ultra-high-speed data transmission at 100 Gbaud and 210 Gbit s⁻¹. **a**, Binary signals at an ultra-high symbol rate of 100 Gbaud, as well as multi-level signals at 70 Gbaud, are generated from a CMOS circuit and amplified to a peak-to-peak voltage of 2.5 V to drive the integrated modulator. **b**, Constellation diagram and reconstructed eye diagram of data transmission at 100 Gbaud. The relatively high BER of 7.8×10^{-5} is limited by the electrical signal quality at this ultra-high speed. **c**, **d**, Multi-level data modulation at a symbol rate of 70 Gbaud. The four-level (**c**) and eight-level (**d**) ASK signals enable even higher data-transmission rates of 140 Gbit s⁻¹ and 210 Gbit s⁻¹, respectively. The measured BERs are 2.1×10^{-5} and 1.5×10^{-2} , respectively. Scale bars, 5 ps.

signal yields error-free performance within the 1.1×10^6 captured signal bits: that is, BER $< 1 \times 10^{-6}$ (Fig. 2c). In this case, the electrical energy dissipation within our modulator is 0.37 fJ bit⁻¹ (see Methods). The system can also operate at a further reduced driving voltage of $V_{pp} = 60$ mV, with a BER of 3.4×10^{-3} (Fig. 2d). In this case, the electrical energy dissipation of the modulator is further reduced to 37 aJ bit⁻¹. We note that the overall energy consumption of the complete data-transmission system is dominated by off-chip components, including CMOS DAC, laser, receiver set-up and analogue-to-digital converters. It is important to consider the whole system when assessing the overall system energy requirements.

The high electro-optic bandwidth and excellent signal fidelity of our modulator allow for data transmission at even higher rates, currently up to 210 Gbit s⁻¹. To achieve this, we amplify the electrical signals from the CMOS DAC to a V_{pp} of about 2.5 V (Fig. 3a). We first test our modulator at an ultra-high symbol rate of 10^{11} symbols per second (100 Gbaud) (Fig. 3b; see Methods)³⁰. The BER of 7.8×10^{-5} in this case is limited by distortion from the electrical source at high frequencies: the electrical BER at 100 Gbaud without any electrical-to-optical-to-electrical conversion is 3.6×10^{-5} (see Methods). We then use multi-level modulation formats at 70 Gbaud to further increase the data rates and to interrogate the signal quality (that is, signal-to-noise ratio, SNR) of our modulator. Using four-level amplitude shift keying (4-ASK)—that is, encoding two bits in each symbol—we can achieve a data-transmission rate of 140 Gbit s⁻¹, with a low BER of 2.1×10^{-5} . In the case of 8-ASK (three bits per symbol), the modulator transmits a total data rate of 210 Gbit s⁻¹. The measured BER of 1.5×10^{-2} is within the tolerance of forward error correction with a 20% overhead (tolerable BER = 1.9×10^{-2}). For the 8-ASK modulation, the electrical energy dissipation of our modulator is 14 fJ bit⁻¹ (see Methods). The high SNR demonstrated here results from a combination of high extinction ratio, low optical loss, linear electro-optic response and the absence of modulation-induced absorption. This further indicates that our devices have maintained the excellent signal fidelity of conventional LN modulators.

We show that the integrated LN platform offers greatly improved overall performance (voltage, bandwidth and optical loss) over traditional LN modulators and other material platforms. By reducing the device

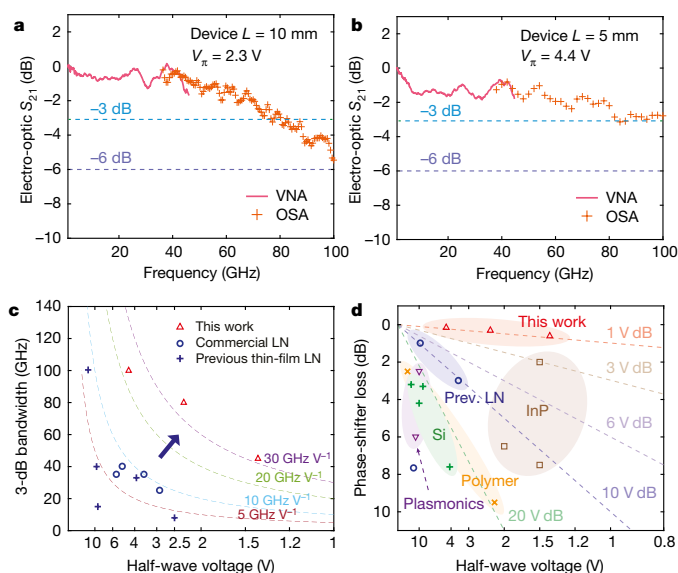


Fig. 4 | Towards ultimate modulator performance. **a**, **b**, Electro-optic responses of a 10-mm (**a**) and a 5-mm (**b**) device, showing ultra-high 3-dB bandwidths of 80 GHz and 100 GHz, respectively. The low-frequency (red line) and high-frequency (orange crosses) measurements are performed using a vector network analyser (VNA) and an optical spectrum analyser (OSA), respectively. **c**, Comparison of key modulator figure of merit (BW/V_{π} , ratio between 3-dB electro-optic bandwidth BW and half-wave voltage V_{π}) for this work, for state-of-the-art commercial LN modulators and for previous thin-film LN modulators, showing much higher BW/V_{π} values of about 30 GHz V⁻¹ for this work. The dotted lines correspond to constant values of BW/V_{π} . **d**, In comparison with other material platforms with high bandwidths, the integrated LN modulators show much lower V_{π} and much lower on-chip optical loss at the same time. The optical losses in the active modulation regions are used for fair comparison between different platforms. Detailed references for the data shown in **c** and **d** can be found in the Methods.

length to 10 mm and 5 mm, we further expand the 3-dB electro-optic bandwidths to 80 GHz and 100 GHz, respectively (Fig. 4a, b), which are measured using an optical spectrum analyser (see Methods)³¹. The measured V_{π} values for these devices are 2.3 V and 4.4 V, respectively (see Methods). Further optimizing the microwave losses of the transmission lines could allow for electro-optic bandwidths of > 100 GHz while maintaining a CMOS-level voltage. Nevertheless, our modulators already represent better voltage–bandwidth performance (about 30 GHz V⁻¹) than commercial LN modulators and previously reported thin-film LN devices, as illustrated in Fig. 4c. The ultra-high bandwidth of these modulators could allow for data operation beyond 200 Gbaud. Furthermore, in contrast to other high-speed modulator platforms in which the optical materials or modulation mechanisms are often inherently lossy, our integrated LN platform breaks the traditional trade-off between modulation voltage and propagation loss, allowing for low optical loss and low V_{π} at the same time (Fig. 4d). Given the voltage–loss performance of our platform, it should be possible to use even longer devices to further reduce the V_{π} to well below 1 V while maintaining an on-chip insertion loss of below 1 dB.

The results presented here show that ultra-high-performance integrated LN modulators have the properties desired for future high-bandwidth and low-power-consumption data communications. By combining the excellent signal fidelity with advanced in-phase/quadrature (I/Q) modulator designs, a single integrated LN modulator could transmit more than 1 Tb s⁻¹ of data per polarization, using, for example, 64-quadrature amplitude modulation (64-QAM) at 200 Gbaud. Using meandering optical waveguides and microwave transmission lines could lead to modulators with even smaller footprint while maintaining CMOS-compatible voltages. This could open opportunities for direct optoelectronic integration of switching components with

application-specific integrated circuits². Furthermore, the low optical and microwave losses, linear electro-optic response, scalability and the ability to be integrated with other photonic components (such as filters and delay lines) could inspire a new generation of active integrated optoelectronic circuits that can be reconfigured on a picosecond timescale using attojoules of electrical energy. These applications include large-scale gigahertz switching networks for quantum photonics⁵, radio signal processing in the optical domain for microwave photonics⁴, self-aware optical networks¹⁰, non-reciprocal devices⁸, topological photonic circuits³² and photonic neural networks¹¹.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0551-y>.

Received: 27 April 2018; Accepted: 25 July 2018;

Published online 24 September 2018.

1. Reed, G. T., Mashanovich, G., Gardes, F. Y. & Thomson, D. J. Silicon optical modulators. *Nat. Photon.* **4**, 518–526 (2010).
2. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
3. Fortier, T. M. et al. Generation of ultrastable microwaves via optical frequency division. *Nat. Photon.* **5**, 425–429 (2011).
4. Ghelfi, P. et al. A fully photonics-based coherent radar system. *Nature* **507**, 341–345 (2014).
5. O'Brien, J. L. Optical quantum computing. *Science* **318**, 1567–1570 (2007).
6. Kues, M. et al. On-chip generation of high-dimensional entangled quantum states and their coherent control. *Nature* **546**, 622–626 (2017).
7. Yu, Z. & Fan, S. Complete optical isolation created by indirect interband photonic transitions. *Nat. Photon.* **3**, 91–94 (2009).
8. Tzuang, L. D., Fang, K., Nussenzeig, P., Fan, S. & Lipson, M. Non-reciprocal phase shift induced by an effective magnetic flux for light. *Nat. Photon.* **8**, 701–705 (2014).
9. Wooten, E. L. et al. A review of lithium niobate modulators for fiber-optic communications systems. *IEEE J. Sel. Top. Quantum Electron.* **6**, 69–82 (2000).
10. Miller, D. A. B. Sorting out light. *Science* **347**, 1423–1424 (2015).
11. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
12. Xu, Q., Schmidt, B., Pradhan, S. & Lipson, M. Micrometre-scale silicon electro-optic modulator. *Nature* **435**, 325–327 (2005).
13. Sun, C. et al. Single-chip microprocessor that communicates directly using light. *Nature* **528**, 534–538 (2015).
14. Ogiso, Y. et al. Over 67 GHz bandwidth and 1.5 V InP-based optical IQ modulator with n-i-p-n heterostructure. *J. Lightwave Technol.* **35**, 1450–1455 (2017).
15. Aoki, M. et al. InGaAs/InGaAsP MQW electroabsorption modulator integrated with a DFB laser fabricated by band-gap energy control selective area MOCVD. *IEEE J. Quantum Electron.* **29**, 2088–2096 (1993).
16. Koeber, S. et al. Femtojoule electro-optic modulation using a silicon-organic hybrid device. *Light Sci. Appl.* **4**, e255 (2015).
17. Lee, M. et al. Broadband modulation of light by using an electro-optic polymer. *Science* **298**, 1401–1403 (2002).
18. Haffner, C. et al. Low-loss plasmon-assisted electro-optic modulator. *Nature* **556**, 483–486 (2018).
19. Boyd, R. W. *Nonlinear Optics* (Academic, Cambridge, 2003).
20. Janner, D., Tulli, D., García-Granda, M., Belmonte, M. & Pruneri, V. Micro-structured integrated electro-optic LiNbO₃ modulators. *Laser Photonics Rev.* **3**, 301–313 (2009).
21. Schmidt, R. V. & Kaminow, I. P. Metal-diffused optical waveguides in LiNbO₃. *Appl. Phys. Lett.* **25**, 458–460 (1974).
22. Poberaj, G., Hu, H., Sohler, W. & Günter, P. Lithium niobate on insulator (LNOI) for micro-photonic devices. *Laser Photonics Rev.* **6**, 488–503 (2012).
23. Liang, H., Luo, R., He, Y., Jiang, H. & Lin, Q. High-quality lithium niobate photonic crystal nanocavities. *Optica* **4**, 1251–1258 (2017).
24. Wang, J. et al. High-Q lithium niobate microdisk resonators on a chip for efficient electro-optic modulation. *Opt. Express* **23**, 23072–23078 (2015).
25. Wang, C., Zhang, M., Stern, B., Lipson, M. & Lončar, M. Nanophotonic lithium niobate electro-optic modulators. *Opt. Express* **26**, 1547–1555 (2018).
26. Rao, A. et al. High-performance and linear thin-film lithium niobate Mach-Zehnder modulators on silicon up to 50 GHz. *Opt. Lett.* **41**, 5700–5703 (2016).
27. Chen, L., Xu, Q., Wood, M. G. & Reano, R. M. Hybrid silicon and lithium niobate electro-optical ring modulator. *Optica* **1**, 112–118 (2014).
28. Zhang, M., Wang, C., Cheng, R., Shams-Ansari, A. & Lončar, M. Monolithic ultra-high-Q lithium niobate microring resonator. *Optica* **4**, 1536–1537 (2017).
29. Weigel, P. O. et al. Hybrid silicon photonic-lithium niobate electro-optic Mach-Zehnder modulator beyond 100 GHz. Preprint at <https://arxiv.org/abs/1803.10365> (2018).
30. Chen, X. et al. All-electronic 100-GHz bandwidth digital-to-analog converter generating PAM signals up to 190 Gbaud. *J. Lightwave Technol.* **35**, 411–417 (2017).
31. Chen, X. et al. Characterization of electro-optic bandwidth of ultra-high speed modulators. In *2017 Optical Fiber Communications Conference and Exhibition 1–3* (2017); <https://doi.org/10.1364/OFC.2017.Tu2H.7>
32. Yuan, L., Xiao, M., Lin, Q. & Fan, S. Synthetic space with arbitrary dimensions in a few rings undergoing dynamic modulation. *Phys. Rev. B* **97**, 104105 (2018).

Acknowledgements We thank J. Khan for discussions on the LN platform, H. Majedi for help with the equipment, and C. Reimer, S. Bogdanović, L. Shao and B. Desiatov for feedback on the manuscript. This work is supported in part by the National Science Foundation (NSF) (ECCS1609549, ECCS-1740296 E2CDA and DMR-1231319) and by Harvard University Office of Technology Development (Physical Sciences and Engineering Accelerator Award). Device fabrication is performed at the Harvard University Center for Nanoscale Systems, a member of the National Nanotechnology Coordinated Infrastructure Network, which is supported by the NSF under ECCS award no. 1541959.

Reviewer information Nature thanks M. Hochberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.W., M.Z., X.C., P.W. and M.L. conceived the experiment. C.W., M.Z. and A.S. fabricated the devices. M.Z. and M.B. performed numerical simulations. C.W., M.Z., X.C. and S.C. carried out the device characterization. C.W. wrote the manuscript with contribution from all authors. P.W. and M.L. supervised the project.

Competing interests C.W., M.Z. and M.L. are involved in developing lithium niobate technologies at HyperLight Corporation.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0551-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Device fabrication. Devices are fabricated from a commercial x-cut LN-on-insulator wafer (NANOLN), in which a 600-nm device layer sits on top of a SiO₂/Si-stack substrate. We use electron-beam lithography (EBL) and Ar⁺-based reactive ion etching to define optical waveguides and MZIs in thin-film LN, using a similar process as described in our previous work²⁸. A 1.5- μm -thick polymethyl methacrylate (PMMA) EBL resist is spin-coated and exposed, again using EBL with alignment, to define the microwave transmission line patterns. The PMMA resist is used for a lift-off process to produce the 1.1- μm -thick gold strip lines. High-precision alignment between the two layers is required to prevent excessive propagation loss from the metal electrodes. The structures are then cladded with an 800-nm-thick SiO₂ layer by plasma-enhanced chemical vapour deposition. Finally, the chip edges are diced and polished to improve the fibre-to-chip coupling.

Electro-optic characterization. Electro-optic characterization is performed in the telecommunications C-band using a tunable-wavelength laser source (Santec TSL-510). A three-paddle polarization controller is used to ensure transverse-electric mode excitation. Light is butt-coupled into and out of the chip under test using tapered lensed fibres, with a coupling loss of about 5 dB per facet. We have previously done extensive studies on the dependence of waveguide propagation loss on the waveguide geometry²⁸. In the current work, we use a waveguide top width of 800 nm to ensure single-mode operation. The corresponding propagation loss is approximately 0.2 dB cm⁻¹, higher than the best value reported in our previous work²⁸ owing to more sidewall scattering losses in the narrower waveguide. This results in an overall on-chip phase-shifter insertion loss of <0.5 dB, as estimated by comparing fibre-to-fibre transmission signals for our modulators (total waveguide lengths >3 cm) and bare waveguides with short total lengths. Within the measurement uncertainty (± 0.5 dB), we did not observe measurable loss difference for the two types of devices, thus extracting an upper bound of on-chip losses of 0.5 dB. For our V_{π} measurement, a triangular voltage signal ($V_{pp} = 5$ V) is used to drive the modulator while the optical transmission signal is monitored in real time. A pair of high-speed microwave probes (GGB) is used to deliver the modulation signal to the input port of the transmission line, and to terminate the output of the transmission line with a 50- Ω load. The electro-optic response below 45 GHz is measured with a vector network analyser in a set-up similar to that used in our previous work²⁵.

Extended Data Fig. 1 shows the half-wave voltage measurements for three devices characterized, with active device lengths of 20 mm, 10 mm and 5 mm. The inset of Extended Data Fig. 1a shows the optical transmission on a logarithmic scale, indicating a measured device extinction ratio of 30 dB. The high experimental extinction ratio also implies a high fabrication quality for the Y-junctions, with a nearly ideal 50:50 split ratio. The extinction ratio could be further improved by using directional couplers. The measured V_{π} values for the 10-mm and 5-mm devices are 2.3 V and 4.4 V, corresponding to voltage-length products of 2.3 V cm and 2.2 V cm, respectively.

Ultra-high-speed device characterization. The electro-optic response from 35 GHz to 100 GHz is tested with an optical spectrum analyser (Extended Data Fig. 2a)³¹. A sinusoidal signal (f_i) from a high-speed synthesizer (up to 50 GHz) is used to drive a commercial LN Mach-Zehnder modulator, which has a 3-dB bandwidth of about 35 GHz, with the frequency response gradually rolling off thereafter. The modulator is biased at the transmission null to suppress the carrier frequency, resulting in an output optical signal with two sidebands separated by twice the input frequency ($2f_i$). A 100-GHz photodetector is used to beat the two sidebands and generate a microwave signal up to 100 GHz, which is subsequently used to drive the modulator under test. The modulator optical response is measured by monitoring the sideband power in the optical spectrum analyser. We calibrate the synthesizer output power, the electro-optic response of the commercial Mach-Zehnder modulator and the 100-GHz photodetector response over the frequency range of interest to ensure an accurate characterization of our devices. Note that the actual measurement is performed using a pair of 67-GHz probes (one for microwave signal modulation and the other for 50- Ω termination). The frequency response of the probes is not de-embedded due to the lack of manufacturer data. Therefore, the bandwidth reported here is a lower bound for our devices.

The measurement set-up for high-speed data modulation is shown in Extended Data Fig. 2b. Electrical signals up to 70 Gbaud are directly generated from a CMOS DAC circuit (Socionext OOLA DAC, 3-dB analogue bandwidth 15 GHz, 13-dB analogue bandwidth 35 GHz). The transmitter uses a digital pulse-shaping filter to limit the signal to its Nyquist bandwidth (for example, 35 GHz electrical bandwidth for a 70-Gbaud signal). The electrical signals from the DAC have a peak-to-peak voltage of ~ 0.2 V and can be further amplified or attenuated before being used to drive the modulator under test. The modulator is biased at the transmission null point to generate binary phase-shift keying, 4-ASK and 8-ASK³³. The output optical signal is mixed with a local oscillator in an optical hybrid and is sent into a single-polarization phase-diversity coherent receiver with a pair of balanced 45-GHz photodetectors to extract the in-phase and quadrature components of the modulated light field. Digitized data are collected using an 80-GSa/s real-time oscilloscope

with an analogue bandwidth of 63 GHz (Keysight DSOZ634A). The collected data are post-processed using least-mean-square adaptive filters to equalize the optical channel response, and to generate the constellation diagrams. The eye diagrams are plotted by up-sampling the real part of the recovered constellations using an interpolation filter for better visualization. At 100 Gbaud, the electrical signal is generated by up-conversion and interleaving two 35-GHz electrical signals³⁰. Extended Data Fig. 3 shows the electrical eye diagram at 100 Gbaud, with a BER of 3.6×10^{-5} , limited by the signal distortion from the electronic circuit. This electrical BER floor translates directly into the BER floor of the optically modulated signal at 100 Gbaud.

For optical signal-to-noise ratio (OSNR) measurement, the modulator output optical signal is attenuated to different levels before being amplified, which produces the desired OSNR values, before entering the coherent receiver. The measured OSNR curves are plotted in Extended Data Fig. 4.

Microwave transmission line and velocity matching. Extended Data Fig. 5a, b shows the cross-sectional schematics of the integrated LN modulator and a conventional LN modulator. Extended Data Fig. 5c, d shows the numerically simulated microwave and optical field distributions. The simulations are performed using the finite element method (COMSOL Multiphysics). Owing to the high optical confinement (Extended Data Fig. 5d) in the thin-film LN platform, we can design the metal electrodes to be placed close to the waveguides (spacing <2.5 μm between metal and waveguide edge) without substantially increasing the optical losses (simulated metal absorption loss <0.03 dB cm⁻¹). We design the waveguide width, ridge height and metal gap to achieve a much higher electro-optic efficiency than conventional LN modulators.

More importantly, the SiO₂/Si-stack substrate in our platform can be independently designed as a microwave dielectric to realize optimal microwave-optical group velocity matching without sacrificing electro-optic overlap. In conventional LN modulators, velocity matching is a non-trivial task because of the large discrepancy between the dielectric constants of LN at microwave ($\epsilon_{\text{RF}} = 28$) and optical ($\epsilon_{\text{opt}} = 5$) frequencies. As a result, a buffer SiO₂ layer is used to overcome the large group-velocity mismatch between microwave and optical signals (Extended Data Fig. 5b), which further reduces the already suboptimal electro-optic efficiency. In our platform, we design the thickness of the buried SiO₂ layer and the material of the handle substrate (Si in this case) to achieve velocity matching without sacrificing electro-optic overlap. We design the coplanar waveguide signal linewidth and the substrate SiO₂ thickness such that the group refractive indices for microwave and optics are both approximately 2.2 (Extended Data Fig. 5e), and that the transmission line has an impedance near 50 Ω . The final device has a LN slab thickness of 300 nm, a buried SiO₂ layer thickness of 4.7 μm and a Si substrate thickness of 500 μm .

Modulator energy consumption. Because our modulator uses a transmission line configuration with a 50- Ω load, the electrical energy per bit dissipated in the modulator can be estimated as $W_e = V_{\text{rms}}^2 / (BR)$, where V_{rms} is the root-mean-square drive voltage, B is the bit-rate and R is the driver impedance¹⁶.

For direct CMOS modulation at 70 Gbit s⁻¹ (Fig. 2c, d), the electrical root-mean-square voltage $V_{\text{rms}} = 36$ mV, resulting in a low electrical energy consumption of 0.37 fJ bit⁻¹. For the 210 Gbit s⁻¹ data modulation in Fig. 3d, the electrical V_{rms} is 360 mV, resulting in an electrical energy consumption of 14 fJ bit⁻¹.

The energy consumption of the entire data-transmission system also includes the power consumption of CMOS DAC, laser, optical amplifiers and analogue-to-digital converter.

Modulator figures of merit comparison. Extended Data Table 1 lists the detailed numbers (V_{π} values, 3-dB electro-optic bandwidths and phase-shifter losses) and references of previously demonstrated thin-film LN modulators^{25,26,29,34,35}, commercial LN modulators^{36–38}, and modulators based on silicon^{39–42}, InP^{14,43,44}, polymers^{45,46} and plasmonics^{18,47,48}. These numbers are used in Fig. 4c, d. The loss numbers plotted in Fig. 4d consider only the loss of the active modulation regions, excluding the coupling and splitter losses, to provide a comparison of modulation-induced optical absorption between different material platforms and modulation schemes. In practical applications, other important performance merits, including footprint, extinction ratio, modulator linearity, thermal stability and power handling, should also be considered. Also, the comparison here is only on the device level. In practical data-transmission settings, the performance of other components, such as the driving circuits, also needs to be considered.

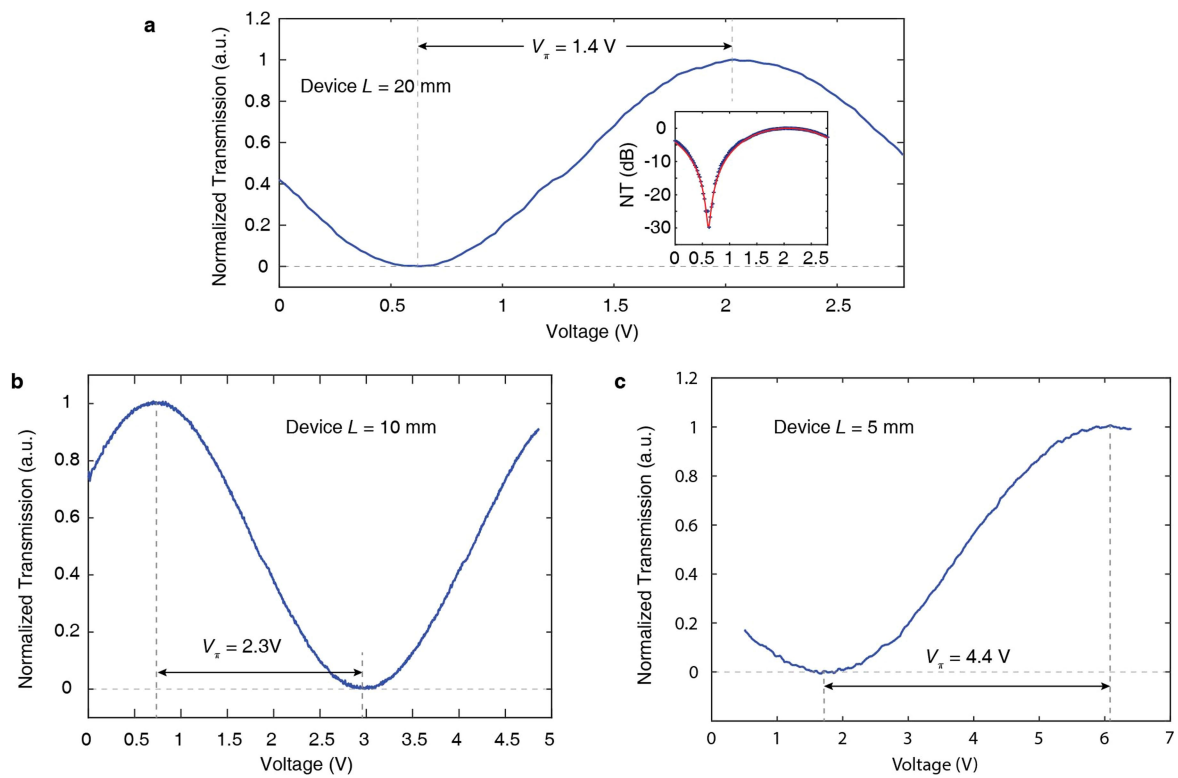
Data availability

The data sets generated and/or analysed during the current study are available from the corresponding authors on reasonable request.

33. Winzer, P. J. & Essiambre, R. J. Advanced optical modulation formats. *Proc. IEEE* **94**, 952–985 (2006).

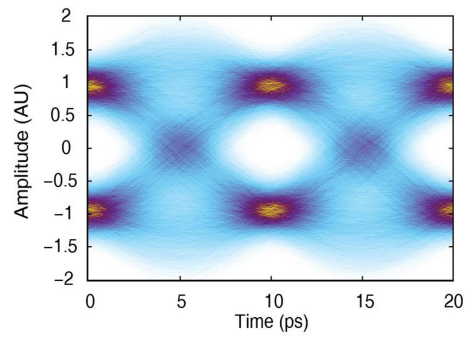
34. Mercante, A. J. et al. 110 GHz CMOS compatible thin film LiNbO₃ modulator on silicon. *Opt. Express* **24**, 15590–15595 (2016).

35. Jin, S., Xu, L., Zhang, H. & Li, Y. LiNbO₃ thin-film modulators using silicon nitride surface ridge waveguides. *IEEE Photonics Technol. Lett.* **28**, 736–739 (2016).
36. 100G/400G LN Modulator. <http://www.fujitsu.com/jp/group/foc/en/products/optical-devices/100gln/>
37. Eospace 2017 Advanced Products. <http://eospace.com/pdf/EOSPACEbriefProductInfo2017.pdf>
38. 40 GHz or 40 Gb/s Lithium Niobate Modulators. https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=3948
39. Dong, P. et al. Monolithic silicon photonic integrated circuits for compact 100+ Gb/s coherent optical receivers and transmitters. *IEEE J. Sel. Top. Quantum Electron.* **20**, 150–157 (2014).
40. Thomson, D. J. et al. 50-Gb/s silicon optical modulator. *IEEE Photonics Technol. Lett.* **24**, 234–236 (2012).
41. Streshinsky, M. et al. Low power 50 Gb/s silicon traveling wave Mach–Zehnder modulator near 1300 nm. *Opt. Express* **21**, 30350–30357 (2013).
42. Azadeh, S. S. et al. Low V silicon photonics modulators with highly linear epitaxially grown phase shifters. *Opt. Express* **23**, 23526–23550 (2015).
43. Rouvalis, E. Indium phosphide based IQ-modulators for coherent pluggable optical transceivers. In *2015 IEEE Compound Semiconductor Integrated Circuit Symposium* 1–4 (2015); <https://doi.org/10.1109/CSICS.2015.7314513>
44. Letal, G. et al. Low loss InP C-band IQ modulator with 40 GHz bandwidth and 1.5 V V_π. In *2015 Optical Fiber Communications Conference and Exhibition* 1–3 (2015); <https://doi.org/10.1364/OFC.2015.Th4E.3>
45. Wolf, S. et al. Coherent modulation up to 100 GBd 16QAM using silicon-organic hybrid (SOH) devices. *Opt. Express* **26**, 220–232 (2018).
46. Alloatti, L. et al. 100 GHz silicon–organic hybrid modulator. *Light Sci. Appl.* **3**, e173 (2014).
47. Ayata, M. et al. High-speed plasmonic modulator in a single metal layer. *Science* **358**, 630–632 (2017).
48. Haffner, C. et al. All-plasmonic Mach–Zehnder modulator enabling optical high-speed communication at the microscale. *Nat. Photon.* **9**, 525–528 (2015).

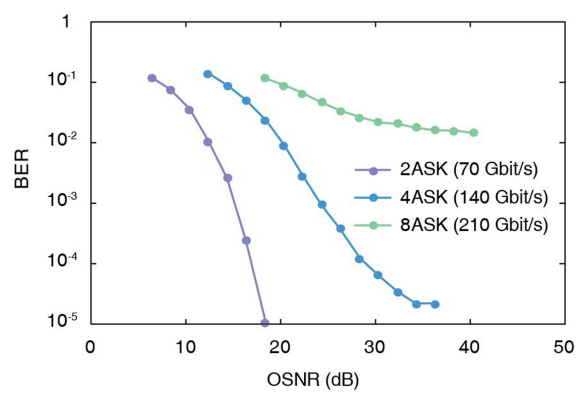


Extended Data Fig. 1 | Half-wave voltages of devices with different active lengths. a–c, Normalized optical transmission of the 20-mm (a), 10-mm (b) and 5-mm (c) device as a function of the applied voltage,

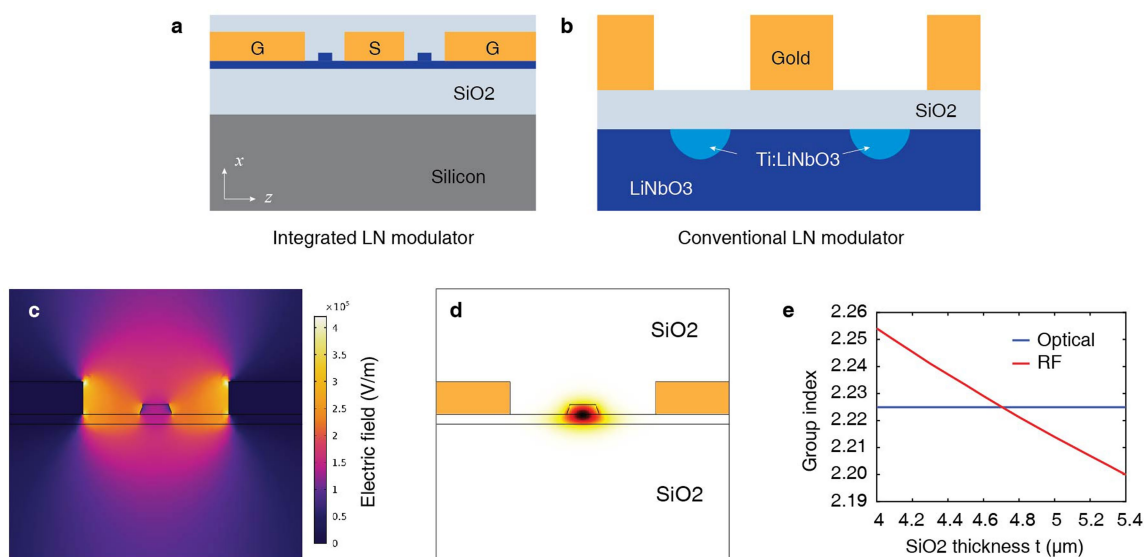
showing half-wave voltages of 1.4 V, 2.3 V and 4.4 V, respectively. The inset of a shows the measured normalized transmission (NT) on a logarithmic scale, revealing an extinction ratio of 30 dB.



Extended Data Fig. 3 | Electrical eye diagram at 100 Gbaud. The measured electrical BER is 3.6×10^{-5} , limited by the signal distortion from the electronic circuit.



Extended Data Fig. 4 | OSNR measurements. BER versus OSNR for the three modulation schemes at 70 Gbaud.



Extended Data Fig. 5 | Comparison of integrated and conventional LN modulators. **a, b**, Schematics of the cross-sections of thin-film (**a**) and conventional (**b**) LN modulators. Our thin-film modulator (**a**) has an oxide layer underneath the device layer, so that velocity matching can be achieved while maximum electro-optic efficiency is maintained. A conventional modulator (**b**) also uses a buffer oxide layer for velocity matching, but on top of LN which further compromises the electro-optic

overlap. **c, d**, Numerically simulated microwave (**c**) and optical (**d**) field distributions (both shown in E_z components) in the cross-section of the thin-film modulator. For microwave simulations, the electric-field values are obtained when a voltage of 1 V is applied across the two electrodes. **e**, Group refractive indices for both optical and microwave signals as a function of the buried oxide thickness. Velocity matching can be achieved with an oxide thickness of about 4,700 nm.

Extended Data Table 1 | Comparison of modulator voltage, bandwidth and loss performance

Type	Half-wave voltage	Voltage-length product	3-dB electro-optic bandwidth	Phase-shifter loss	Reference
Thin-film LN	1.4 V	2.8 V·cm	45 GHz	0.4 dB	This work
Thin-film LN	2.3 V	2.3 V·cm	80 GHz	0.2 dB	This work
Thin-film LN	4.4 V	2.2 V·cm	100 GHz	0.1 dB	This work
Thin-film LN	4 V	3.1 V·cm	33 GHz	N/A	[26]
Thin-film LN	9.4 V	9.4 V·cm	40 GHz	1 dB	[34]
Thin-film LN	9 V	1.8 V·cm	15 GHz	0.6 dB	[25]
Thin-film LN	2.5 V	3 V·cm	8 GHz	8.4 dB	[35]
Thin-film LN	13 V	6.7 V·cm	100 GHz	7.8 dB	[29]
Commercial LN	3.5 V	>10 V·cm	35 GHz	N/A	Fujitsu, [36]
Commercial LN	4.5 - 4.9 V	>10 V·cm	30 - 40 GHz	4 dB	EO Space, [37]
Commercial LN	2.9 - 3.3 V	>10 V·cm	20 - 25 GHz	4 dB	EO Space, [37]
Commercial LN	5.5 V	>10 V·cm	35 GHz	N/A	Thorlabs, [38]
Silicon	10 V	2.4 V·cm	N/A	4.2 dB	[39]
Silicon	16 V	2.8 V·cm	N/A	3.2 dB	[40]
Silicon	8.5 V	2.6 V·cm	30 GHz	3.3 dB	[41]
Silicon	4.1 V	0.74 V·cm	34 GHz	7.6 dB	[42]
Indium Phosphide	2 V	N/A	40 GHz	6.5 dB	[43]
Indium Phosphide	1.5 V	0.6 V·cm	40 GHz	7.5 dB	[44]
Indium Phosphide	1.5 V	0.54 V·cm	67 GHz	2 dB	[14]
Polymer	2.2 V	0.11 V·cm	100 GHz	9.5 dB	[45]
Polymer	22 V	0.11 V·cm	100 GHz	2.5 dB	[46]
Plasmonics	20 V	N/A	110 GHz	2.5 dB	[18]
Plasmonics	12 V	0.012 V·cm	70 GHz	6 dB	[47]
Plasmonics	10 V	0.006 V·cm	70 GHz	2.5 dB	[48]

Detailed half-wave voltages, 3-dB electro-optic bandwidths and phase-shifter losses of commercial LN modulators, previously demonstrated thin-film LN modulators as well as modulators based on Si, InP, polymers and plasmonics with high bandwidths, are listed.

Asymmetric α -arylation of amino acids

Daniel J. Leonard¹, John W. Ward¹ & Jonathan Clayden^{1*}

Quaternary amino acids, in which the α -carbon that bears the amino and carboxyl groups also carries two carbon substituents, have an important role as modifiers of peptide conformation and bioactivity and as precursors of medicinally important compounds^{1,2}. In contrast to enantioselective alkylation at this α -carbon, for which there are several methods^{3–8}, general enantioselective introduction of an aryl substituent at the α -carbon is synthetically challenging⁹. Nonetheless, the resultant α -aryl amino acids and their derivatives are valuable precursors to bioactive molecules^{10,11}. Here we describe the synthesis of quaternary α -aryl amino acids from enantiopure amino acid precursors by α -arylation without loss of stereochemical integrity. Our approach relies on the temporary formation of a second stereogenic centre in an N' -aryleurea adduct¹² of an imidazolidinone derivative⁶ of the precursor amino acid, and uses readily available enantiopure amino acids both as a precursor and as a source of asymmetry. It avoids the use of valuable transition metals, and enables arylation with electron-rich, electron-poor and heterocyclic substituents. Either enantiomer of the product can be formed from a single amino acid precursor. The method is practical and scalable, and provides the opportunity to produce α -arylated quaternary amino acids in multi-gram quantities.

Among the most practical and widely used methods^{13,14} for the synthesis of α -alkylated amino acids are those that use a readily available chiral amino acid both as a starting material and as a source of chirality, using the principle of ‘self-regeneration of stereocentres’¹⁵. This strategy relies on the diastereoselective formation of an imidazolidinone or oxazolidinone, which creates a new stereogenic centre. The configuration of this stereocentre is retained during the formation of a planar amino acid enolate, and it then directs alkylation of the enolate to form a quaternary stereocentre with control over absolute configuration.

The mechanistically unusual¹⁵ N-to-C aryl migration that occurs in anionic derivatives of ureas was first reported in the construction of stereodefined quaternary centres from configurationally stable organolithiums¹², and it has been used to prepare racemic 5,5-disubstituted hydantoins¹⁶. Stereoselective versions of this hydantoin synthesis using conformational chiral memory¹⁷ or a stoichiometric auxiliary¹⁸ suggested that a practical stereoselective modification of this intramolecular arylation based on imidazolidinone alkylation chemistry might offer a strategy for the synthesis of unavailable enantiopure α -arylated amino acids (Fig. 1).

We therefore explored N' -aryl ureas as a potential intramolecular source of the coupling partner for a corresponding arylation reaction. A versatile synthesis of the N -carbamoylimidazolidinones **3** was required, and our initial synthetic approach is shown in Fig. 2a. Treatment of L-AlaNHMe with pivaldehyde and trifluoroacetic acid formed the *trans* diastereoisomer of the imidazolidinone trifluoroacetate salt with good selectivity¹⁹. In situ chloroformylation with triphosgene in base gave high yields of the N -chloroformylimidazolidinones **1-Ala**, as a 4:1 mixture of the *trans* and *cis* diastereoisomers *trans*-**1-Ala** and *cis*-**1-Ala**. These were readily separated by column chromatography and their relative configurations were established by X-ray crystallography (Fig. 2b) and nuclear Overhauser effect experiments (Supplementary Information).

The minor diastereoisomer *cis*-**1-Ala** acylated N -methylaniline (PhNHMe) cleanly in refluxing dichloromethane to give the urea

cis-**3-Ala-a** in high yield (Fig. 2a, Extended Data Table 1, entry 1). The major *trans* diastereoisomer of **1-Ala** (which characteristically and diagnostically exhibited slow N–CO rotation by NMR; Supplementary Information) was much less reactive. The urea *trans*-**3-Ala-a** was formed only when *trans*-**1-Ala** was activated with potassium iodide²⁰, and a reaction time of 45 h in refluxing CH₂Cl₂ was required for acceptable yields (Fig. 2a, Extended Data Table 1, entries 2–4).

We were now in a position to address the question of the key C–C bond forming step: whether ureas **3-Ala** can undergo the rearrangement we had discovered with other amino acid enolates to provide a means of arylating the amino acid α -centre in a diastereoselective manner. *cis*- and *trans*-**3-Ala-a** were each cooled and treated with base to form an enolate, which was allowed to warm to room temperature. Initial experiments with lithium diisopropylamide (LDA) showed that enolate formation was complete at -78°C (Extended Data Table 2, entry 1), and that warming to room temperature was sufficient to induce 1,4 migration of the phenyl ring to the enolate carbon to yield the C-arylated product imidazolidinone **4-Ala-a** from *trans*-**3** and its enantiomer *ent*-**4-Ala-a** from *cis*-**3** (Extended Data Table 2, entries 2, 3). The best yields were obtained on forming the enolate at 0°C , and even with the milder base potassium bis(trimethylsilyl)amide (KHMDS), **4-Ala-a** was formed in 95% yield from *trans*-**3-Ala** as a single diastereoisomer on a >1-g scale (Extended Data Table 2, entry 4). These conditions (shown as method A in Fig. 2a) were identified as optimal, and a similar yield of the enantiomeric product *ent*-**4-Ala** was obtained under these conditions from *cis*-**3-Ala** (Extended Data Table 2, entry 5). In neither case was any trace of the other diastereoisomer of **4-Ala** detectable in the product by ¹H NMR, and high-performance liquid chromatography on a chiral stationary phase indicated that the product was essentially enantiomerically pure, with an enantiomeric ratio (e.r.) of >99:1.

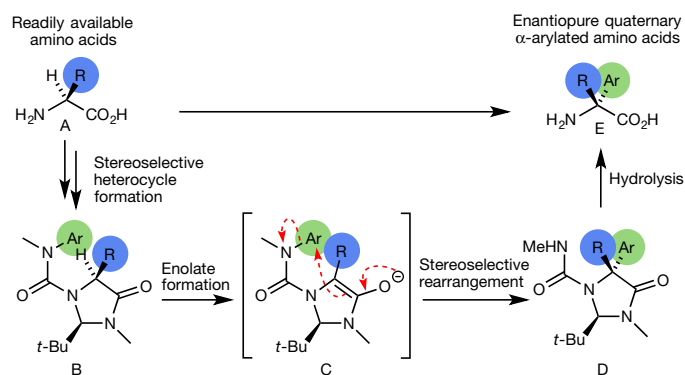


Fig. 1 | Stereoselective arylation of amino acids. Our strategy for stereoselective arylation of amino acids by way of imidazolidinyl ureas is shown. An amino acid (A) is converted diastereoselectively into an imidazolidinone (B) carrying a pendent urea function. Treatment with base forms an enolate (C) in which the aromatic substituent (Ar) of the urea migrates to the rear face of the imidazolidinone, directed by the bulky *tert*-butyl group, as indicated by the red dotted arrows. Hydrolysis of the product (D) provides the quaternary α -aryl amino acid (E).

¹School of Chemistry, University of Bristol, Bristol, UK. *e-mail: j.clayden@bristol.ac.uk

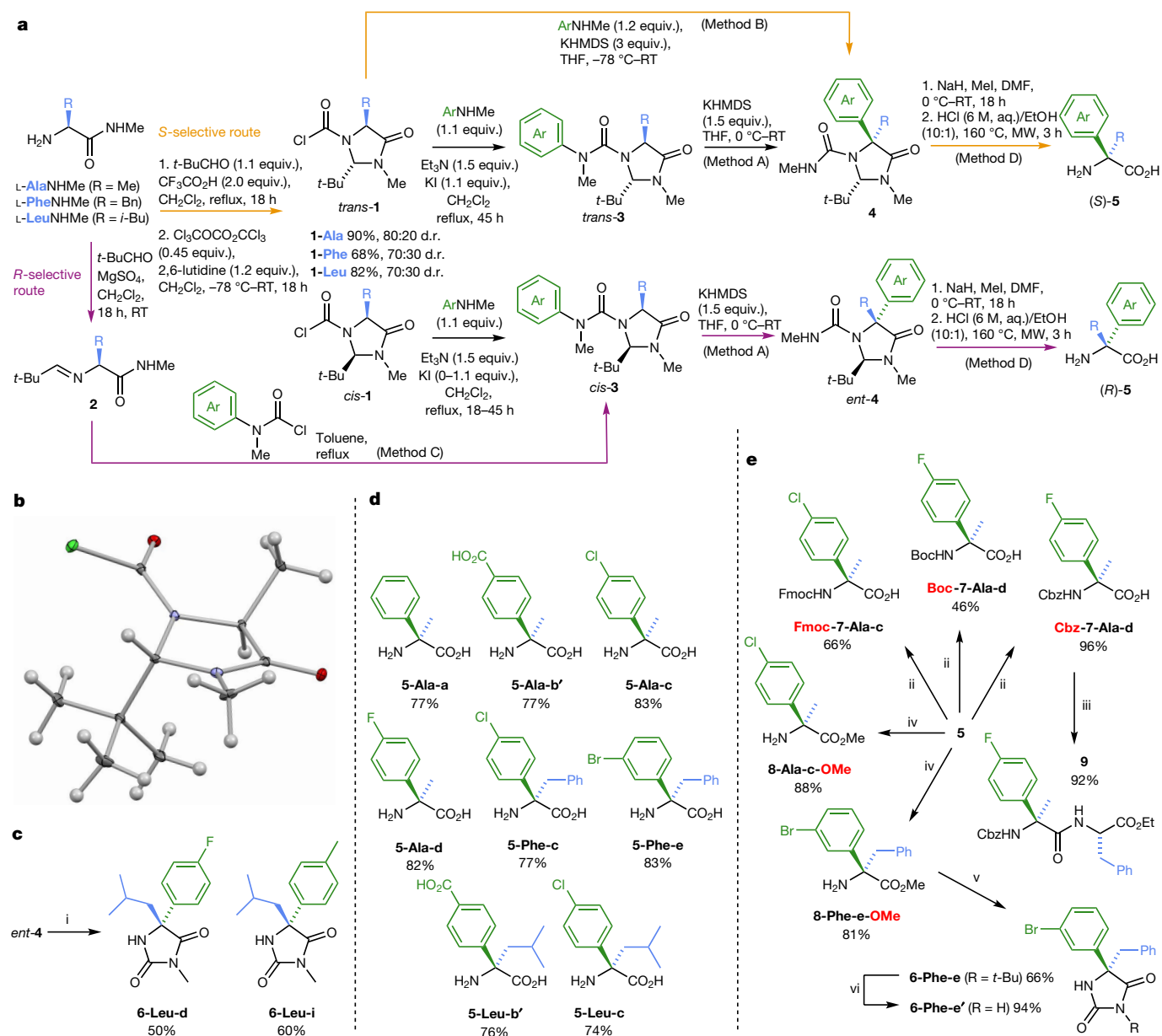


Fig. 2 | Arylation of amino acids by way of imidazolidinone ureas.

a, Synthetic pathways from L-amino acids to quaternary α -arylated amino acids **5** by way of *N*-chloroformylimidazolidinones **1** or imines **2**, *N'*-aryl imidazolinyl ureas **3**, and *C*-aryl imidazolidinones **4**. The sequence shown by the orange arrows starting from **1** constitutes an *S*-selective route to **5** from an L-amino acid, whereas the sequence shown by the purple arrows from **2** constitutes an *R*-selective route from an L-amino acid. DMF, dimethylformamide; MW, microwave; RT, room temperature. **b**, The stereochemistry of *trans*-**1-Ala** is confirmed by X-ray crystallography. **c**, Representative α -arylated hydantoin derivatives formed by hydrolysis of *ent*-**4**. Conditions: i, HCl (6 M, aq.), 130 °C (sealed tube), 18 h. **d**, Yields of

representative α -arylated amino acids **5** formed by the methylation and hydrolysis of **4**. **e**, Derivatization of representative quaternary α -arylated amino acids **5** by *N*-protection, peptide coupling, esterification or hydantoin formation. Conditions: ii, 1. *N*-Methyl-*N*-(trimethylsilyl) trifluoroacetamide, CH_2Cl_2 , reflux, 4 h; 2. CbzOSu or Boc₂O or FmocOSu (OSu, *N*-hydroxysuccinimide), CH_2Cl_2 , RT, 16 h; 3. MeOH, RT, 15 min; iii, 1. K-Oxyma, EDC-HCl (EDC, 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide), *i*-Pr₂NEt, DMF, 0 °C-RT, 15 min; 2. L-Phe-OEt-HCl, *N,N*-diisopropylethylamine, 72 h; iv, Me₃SiCHN₂, benzene/MeOH (4:1), RT, 18 h; v, 1. *t*-BuNCO, CH_2Cl_2 , reflux, 18 h; 2. *t*-BuOK, THF, RT, 18 h; vi, HBr, acetic acid (1:1), 120 °C, 18 h.

Either enantiomer of the product **4-Ala** could be formed from the same L-Ala starting material, simply by the choice of route. However, some work on the synthesis of **3** was still needed for this to become a general method for the arylation of amino acids other than alanine. Two problems remained: first, although *cis*-**3-Phe** was successfully formed from *cis*-**1** in the presence of KI (Extended Data Table 1, entry 5), *cis*-**1** was generally available only in impractically small quantities as it is formed as the minor diastereoisomer in the preceding chloroformylation step. Second, the unreactivity of the major diastereoisomer *trans*-**1** meant that *trans*-**3** could not be formed reliably by this route from amino acids other than alanine: attempted acylations using *trans*-**1-Phe**

were unproductive even when using KI as an activator (Extended Data Table 1, entry 6).

A more robust synthesis of *trans*-**4** was obtained by returning to the easily formed *N*-chloroformylimidazolidinones *trans*-**1** as alternative precursors. Although acylation of a neutral *N*-methylaniline with *trans*-**1** had proved insufficiently general as a way of making **3** (Extended Data Table 1, entry 6), reaction of *trans*-**1-Ala**, *trans*-**1-Phe** or *trans*-**1-Leu** with the anions of a range of *N*-methyl anilines, formed using an excess of KHMDs, not only promoted the acylation of the amine to give *trans*-**3** but also led to deprotonation and rearrangement of **3** to give **4**. Optimized conditions for this one-pot procedure (labelled method B

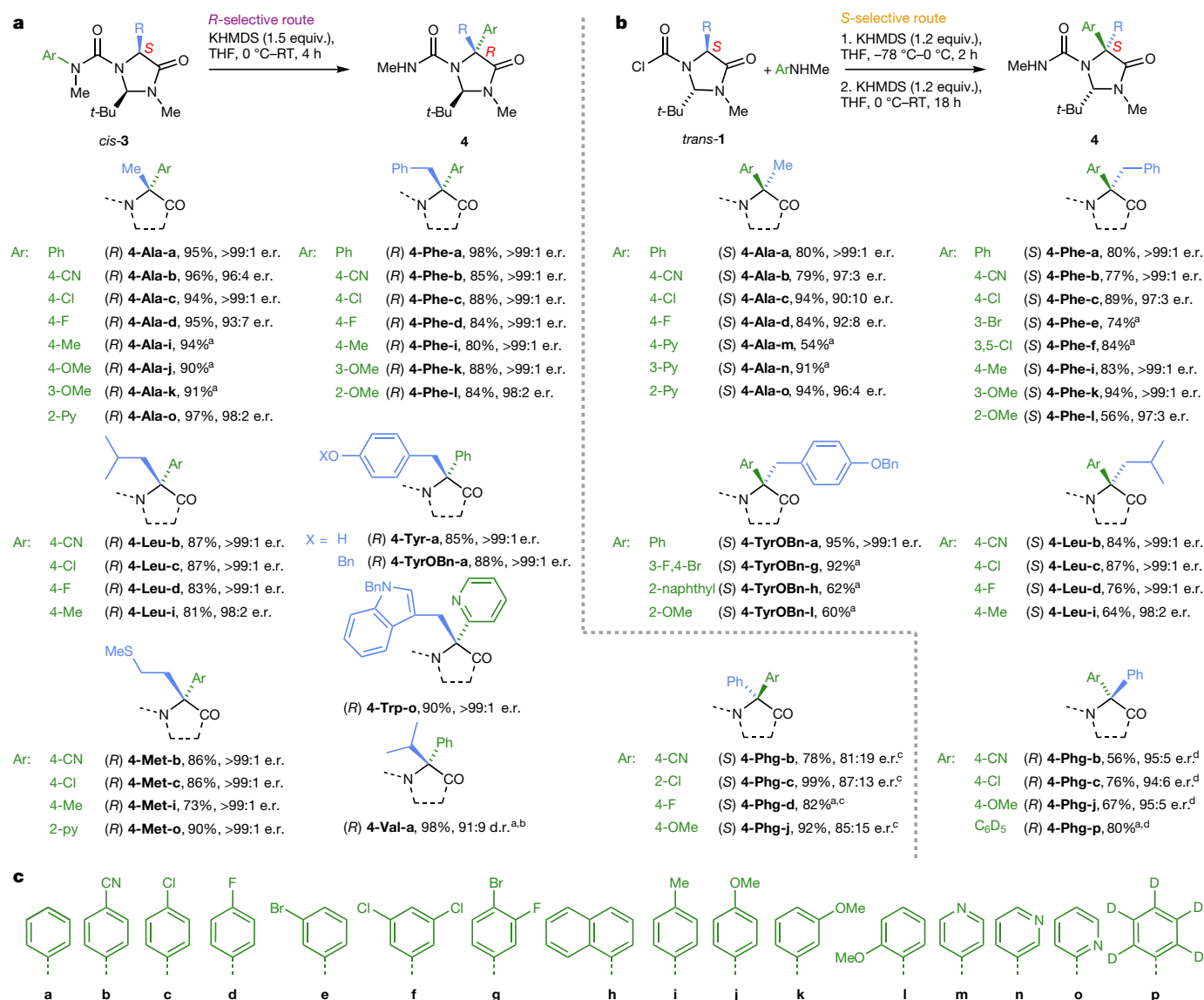


Fig. 3 | Scope of the imidazolidinone arylation: amino acids and migrating groups. **a**, Product structures, yields and e.r. from use of the optimized *R*-selective route via L-amino-acid-derived imidazolidinones *cis*-**3**. Ar indicates either the aryl substituent itself or the substituent(s) on a phenyl ring. **b**, Product structures, yields and e.r. from use of the optimized *S*-selective route via L-amino-acid-derived imidazolidinones *trans*-**1**. **c**, Structures of the aryl substituents introduced by these

methods. ^ae.r. not determined; ^bLiNEt₂ used instead of KHMDS (which gave no product). The product **4-Val-a** contained some of the epimeric imidazolidinone as a result of incompletely diastereoselective rearrangement. ^cD-Phenylglycine was used as starting material, so product has *S* absolute configuration. ^dD-Phenylglycine was used as starting material, so product has *R* absolute configuration.

in Fig. 2a) involved two separate additions of KHMDS. Method B provided an efficient synthesis of an array of products, including **4-Ala**, **4-Phe** and **4-Leu**, which bore a representative selection of substituted aryl rings in high yield and high diastereoselectivity (Supplementary Information).

To explore a similarly efficient route to *ent*-**4** from the same L-amino acids, we turned to an alternative synthesis of *cis*-**3** with complementary diastereoselectivity. It has been shown that, whereas *trans* imidazolidinones are formed at lower temperatures under acidic conditions, diastereoselectivity towards *cis* *N*-acylimidazolidinones can be achieved by acylation of the pivaldimine derivatives of amino acids, probably because of the *cis*-selectivity exhibited by cyclization of the hindered, planar *N*-acyliminium intermediate²¹. We found that urea *cis*-**3-Ala** was indeed formed when the imine **2-Ala** was acylated with *N*-methyl-*N*-phenylcarbamoyl chloride (Fig. 2a, Extended Data Table 1, entries 7, 8). Optimal yields of the pure *cis* diastereoisomer were obtained in refluxing toluene or dichloroethane in the presence of 5 mol% 4-dimethylaminopyridine (entries 10, 11), but with stoichiometric

Et₃N no product was obtained (entry 9). We assume that under these conditions of nucleophilic catalysis, cyclization to the imidazolidinone is reversible, with the rather unreactive carbamoyl chloride selectively acylating the less hindered *cis* diastereoisomer. The method was successfully used to form *cis*-*N*-carbamoylimidazolidinones *cis*-**3-Ala**, *cis*-**3-Phe** and *cis*-**3-Leu** bearing substituted aryl rings by way of their imines **2** (Supplementary Information). These imidazolidinone substrates were subjected to the conditions (method A) previously optimized for *cis*- and *trans*-**3-Ala** to yield the products *ent*-**4**, enantiomeric with those formed from *trans*-**3**.

The *S*-selective and *R*-selective routes highlighted by the orange and purple arrows in Fig. 2a thus provide enantiocomplementary syntheses of the imidazolidinones **4** and *ent*-**4** from the representative L-amino acids L-Ala, L-Phe and L-Leu. These structures are simple derivatives of quaternary amino acids, and were converted into the target α-arylated amino acids **5** by hydrolysis under acidic conditions. Excellent yields of the enantiopure amino acids **5** were obtained by *N*-methylation of the urea function of **4** followed by microwave heating with 6 M HCl

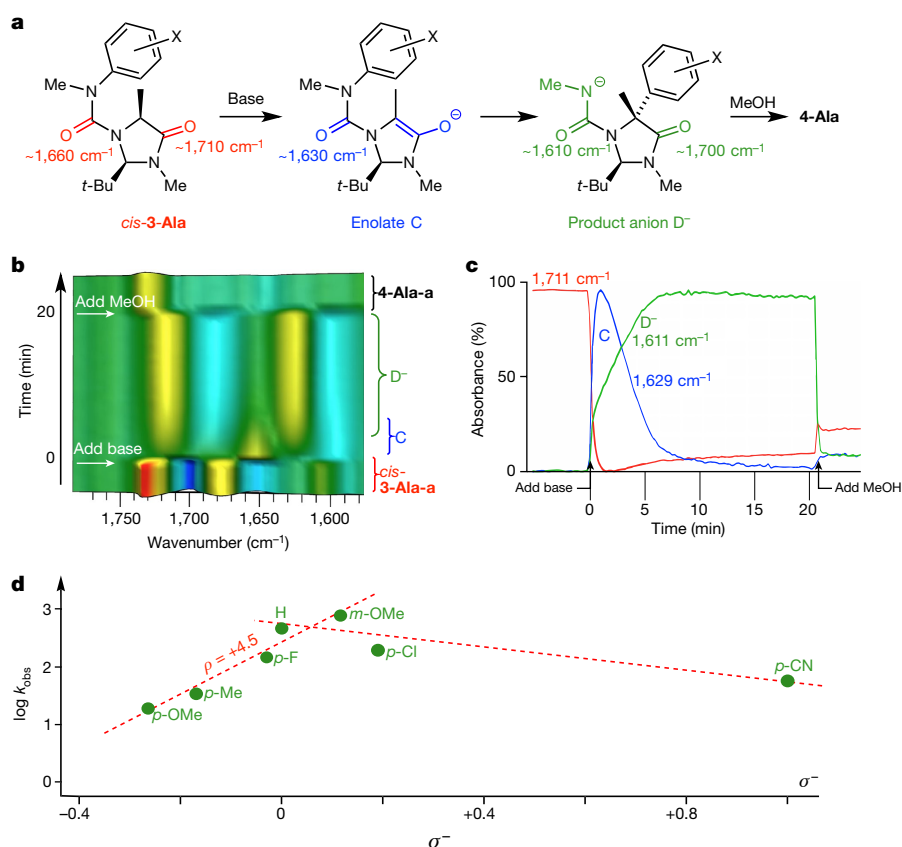


Fig. 4 | Mechanism of the rearrangement.

a, Proposed reaction pathway, with approximate C=O stretching frequencies. **b**, In situ infrared trace (first-derivative plot) of the reaction of *cis*-3-Ala-a, showing diagnostic changes in carbonyl-stretching frequencies. **c**, Plot of absorbance against time for peaks at 1,711 cm⁻¹ (red, starting material), 1,629 cm⁻¹ (blue, enolate (C)) and 1,611 cm⁻¹ (green, product anion (D⁻)). **d**, Hammett plot of log *k*_{obs} against σ^- , consistent with rate-determining rearrangement for electron-rich rings and rate-determining deprotonation for electron-deficient rings. The gradient of the electron-rich domain on the left of the plot, $\rho = +4.5$, is consistent with substantial charge build-up on the aryl substituent during the rearrangement.

(Fig. 2a, method D, Fig. 2d). The *p*-cyano function of 4-Ala-b and 4-Leu-b was hydrolysed under these conditions to give the carboxylated phenylglycine derivatives 5-Ala-b' and 5-Leu-b'. 5-Ala-b' is the mGluR antagonist (S)-M4CPG¹⁰.

Hydrolysis without preliminary *N*-methylation led to competitive formation of the corresponding *N*-methylhydantoin 6 in moderate yield (Fig. 2c) owing to cyclization of the urea onto the newly revealed carboxylic acid. These hydantoin 6 could be hydrolysed cleanly to 5 in a second step, but are nonetheless themselves valuable target structures². A more versatile synthesis of 6 was obtained by treatment of the methyl ester 8-Phe-e-OMe with *tert*-butyl isocyanate to give 6-Phe-e, the *tert*-butyl group being removable to give 6-Phe-e' under acidic conditions (Fig. 2e).

Derivatives of the arylated amino acids that are of value in synthetic procedures such as peptide formation were also formed from 5 (Fig. 2e). Protection of the amino or carboxyl group gave the carboxybenzyl (Cbz)-, *tert*-butoxycarbonyl (Boc)- or fluorenylmethoxycarbonyl (Fmoc)-protected carbamates 7 and the esters 8. Despite the steric hindrance of the quaternary amino acid, dipeptide 9 was formed cleanly on coupling Cbz-7-Ala-d to L-Phe-OEt under standard conditions.

After showing that the base-promoted rearrangement of 3 provides a viable method for the arylation of an initial selection of amino acids, we returned to the synthesis of 4 and *ent*-4 with the aim of extending the scope to the synthesis of other amino acids, and exploring the scope of migrating aryl groups that can be tolerated by the method. The optimal conditions of the *S*-selective route and the *R*-selective route were applied to a range of starting materials that were derived from amino acids, and the successful outcomes of these reactions are summarized in Fig. 3.

Halogenated (c–g) rings, even those bearing bromo substituents, rearranged without evidence of dehalogenation or benzyne formation. Sterically hindered *ortho*-substituted (i) and 1-naphthyl rings (h) also rearranged in good yield. Despite the fact that the rearrangement is formally an intramolecular nucleophilic aromatic substitution (S_NAr) reaction, it shows remarkable tolerance to variations in the aryl

migrating group, with conjugated (h), electron-deficient (b) and electron rich (i–l) rings all taking part in the reaction. All three orientations of a pyridyl ring (m–o) gave rearranged products regiospecifically.

Beyond alanine, phenylalanine and leucine, the functionalized side chains of methionine, tyrosine and tryptophan were tolerated, with arylation of tyrosine being successful even without the protection of its hydroxyl group. Phenylglycine (Phg) was also arylated, enabling the enantioselective synthesis of chiral diaryl glycine derivatives 4-Phg (including the enantioselectively deuterated 4-Phg-p). With phenylglycine, it was necessary to use method B (starting from *trans*-1-Phg) to ensure high enantiomeric ratios, as its acidifying side chain evidently leads to some racemization in the synthesis of *cis*-3-Phg via imine 2-Phg. Arylation of the sterically hindered 3-Val failed with KHMDS, but rearrangement of 3-Val-a to 4-Val-a proceeded in excellent yield with lithium diethylamide, a more powerful and less bulky base. A slight loss in diastereoselectivity was seen in this reaction, possibly due to the more demanding steric requirements of a transition state in which the *tert*-butyl and isopropyl groups are both on the same side of the imidazolidinone ring.

The mechanism by which the enolate of 3 forms 4 is intriguing. The reaction bears some similarity to the Smiles and Truce–Smiles rearrangements^{22,23}, but is distinguished from almost all known examples of these rearrangements by the lack of requirement for an electron-deficient migrating ring. Sensitivity to electronic features may be measured by the Hammett reaction constant ρ , and we explored the kinetics of the reaction by in situ infrared spectroscopy in order to estimate a value of ρ for the rearrangement.

Preliminary studies by infrared spectroscopy using *cis*-3-Ala-a under the optimized conditions for the reaction (1.5 equiv. KHMDS in tetrahydrofuran (THF) at room temperature) revealed no reaction intermediates, which indicates that rearrangement is faster than enolate formation at room temperature. Changing the base to LDA and carrying out the rearrangement at –20 °C decreased the rate of both deprotonation and rearrangement, and revealed an intermediate on the reaction pathway (Fig. 4a, b). This intermediate was identified

as the enolate C (Fig. 4a), on the basis that it has no C=O stretching absorption corresponding to an amide carbonyl group ($1,710\text{ cm}^{-1}$ in *cis*-3-**Ala-a**), but retains the urea ($1,630\text{ cm}^{-1}$) and aromatic ($1,500$ – $1,600\text{ cm}^{-1}$) bands. The rate of decay of this intermediate was identical for both *cis*-3-**Ala-a** and *trans*-3-**Ala-a**, which confirms that it is a common intermediate from both diastereoisomers, and treatment of the isolated product with LDA gave an infrared spectrum identical to that of the species present at the end of the reaction, identifying it as the product anion D^- (Fig. 4a). Confirmation that the reaction is intramolecular was provided by a crossover experiment in which *cis*-3-**Ala-b** was mixed with *cis*-3-**Met-c** (both of which rearrange at comparable rates) and treated with KHMDS. A mass spectrum of the crude reaction mixture showed molecular ions corresponding only to 4-**Ala-b** and 4-**Met-c** (Supplementary Information).

A Hammett plot was constructed by treating a series of imidazolidinones 3-**Ala** that bore a selection of aryl substituents with an excess (5 equiv.) of LDA at -20°C , and the formation of the product anion D^- was monitored using its characteristic infrared bands at around $1,690$ and $1,630\text{ cm}^{-1}$. Under these conditions, the formation of the product from the enolate followed first-order kinetics, and the linear section of a plot of $\ln([\text{D}^-]_\infty - [\text{D}^-])$ against time gave a rate constant k_{obs} for each substrate (Fig. 4c). A Hammett plot of $\log k_{\text{obs}}$ against the substituent constant σ^- is shown in Fig. 4d: the plot shows a downwards bend characteristic of a change in rate-determining step, with enolate formation being rate-limiting for electron-deficient rings (no enolate was detectable by infrared spectroscopy during the rearrangement of 3-**Ala-b** or 3-**Ala-c**). For the electron-rich domain of the plot, the value of ρ is $+4.5$, consistent with substantial build-up of negative charge on the migrating ring during the reaction. This ρ value is nonetheless smaller in magnitude than those of ‘classical’ intermolecular $\text{S}_{\text{N}}\text{Ar}$ reactions^{24,25}, which possibly indicates that the reaction proceeds without the intermediacy of an anionic Meisenheimer complex^{26–28}. Electron-rich substitution patterns are unreactive in such intermolecular substitutions, and we assume that in our system the conformational restriction imposed by the urea linkage²⁹ must enforce attack of the enolate on the ring, irrespective of the inability of the ring to stabilize a negative charge³⁰. As a consequence, the aryl ring behaves as an electrophile, much as alkylating agents do in ‘classical’ reactions of enolates. This use of conformational restriction to induce electrophilic reactivity in electron-rich substituents not only makes generally available this otherwise elusive class of modified amino acids, but also has the potential for wider application in synthesis.

Data availability

Full experimental details and spectroscopic data are provided as Supplementary Information.

Received: 6 March 2018; Accepted: 17 August 2018;

Published online 3 October 2018.

1. Toniolo, C., Crisma, M., Formaggio, F. & Peggion, C. Control of peptide conformation by the Thorpe–Ingold effect (C^α -tetrasubstitution). *Biopolymers* **60**, 396–419 (2001).
2. Meusel, M. & Gütschow, M. Recent developments in hydantoin chemistry. A review. *Org. Prep. Proced. Int.* **36**, 391–443 (2004).
3. Cativiela, C. & Díaz-de-Villegas, M. D. Recent progress on the stereoselective synthesis of acyclic quaternary α -amino acids. *Tetrahedron Asymmetry* **18**, 569–623 (2007).
4. Hashimoto, T. & Maruoka, K. Recent development and application of chiral phase-transfer catalysts. *Chem. Rev.* **107**, 5656–5682 (2007).
5. Schöllkopf, U. Enantioselective synthesis of non-proteinogenic amino acids via metallated *bis*-lactim ethers of 2,5-diketopiperazines. *Tetrahedron* **39**, 2085–2091 (1983).
6. Seebach, D., Sting, A. R. & Hoffmann, M. Self-regeneration of stereocenters (SRS)—applications, limitations, and abandonment of a synthetic principle. *Angew. Chem. Int. Edn Engl.* **35**, 2708–2748 (1996).
7. Kawabata, T. & Fujii, K. Memory of chirality: asymmetric induction based on the dynamic chirality of enolates. *Top. Stereochem.* **23**, 175–205 (2003).
8. Branca, M. et al. Memory of chirality of tertiary aromatic amides: a simple and efficient method for the enantioselective synthesis of quaternary α -amino acids. *J. Am. Chem. Soc.* **131**, 10711–10718 (2009).

9. Shirakawa, S., Yamamoto, K. & Maruoka, K. Phase-transfer-catalyzed asymmetric $\text{S}_{\text{N}}\text{Ar}$ reaction of α -amino acid derivatives with arene chromium complexes. *Angew. Chem. Int. Ed.* **54**, 838–840 (2015).
10. Ma, D. W. Conformationally constrained analogues of L-glutamate as subtype-selective modulators of metabotropic glutamate receptors. *Bioorg. Chem.* **27**, 20–34 (1999).
11. Sonowal, H. et al. Aldose reductase inhibitor increases doxorubicin-sensitivity of colon cancer cells and decreases cardiotoxicity. *Sci. Rep.* **7**, 3182 (2017).
12. Clayden, J., Dufour, J., Grainger, D. M. & Helliwell, M. Substituted diarylmethylamines by stereospecific intramolecular electrophilic arylation of lithiated ureas. *J. Am. Chem. Soc.* **129**, 7488–7489 (2007).
13. Wang, X.-J. et al. Asymmetric synthesis of LFA-1 inhibitor BIRT2584 on metric ton scale. *Org. Process Res. Dev.* **15**, 1185–1191 (2011).
14. Yee, N. K. et al. Practical synthesis of a cell adhesion inhibitor by self-regeneration of stereocenters. *Tetrahedron Asymmetry* **14**, 3495–3501 (2003).
15. Grainger, D. M. et al. The mechanism of the stereospecific intramolecular arylation of lithiated ureas: the role of Li^+ probed by electronic structure calculations, and by NMR and IR spectroscopy. *Eur. J. Org. Chem.* **4**, 731–743 (2012).
16. Atkinson, R. C. et al. Intramolecular arylation of amino acid enolates. *Chem. Commun.* **49**, 9734–9736 (2013).
17. Tomohara, K., Yoshimura, T., Hyakutake, R., Yang, P. & Kawabata, T. Asymmetric α -arylation of amino acid derivatives by Clayden rearrangement of ester enolates via memory of chirality. *J. Am. Chem. Soc.* **135**, 13294–13297 (2013).
18. Atkinson, R. C., Fernández-Nieto, F., Mas Roselló, J. & Clayden, J. Pseudoephedrine-directed asymmetric α -arylation of α -amino acid derivatives. *Angew. Chem. Int. Ed.* **54**, 8961–8965 (2015).
19. Nagib, D. A., Scott, M. E. & MacMillan, D. W. C. Enantioselective α -trifluoromethylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **131**, 10875–10877 (2009).
20. Wakeham, R. J., Taylor, J. E., Bull, S. D., Morris, J. A. & Williams, J. M. J. Iodide as an activating agent for acid chlorides in acylation reactions. *Org. Lett.* **15**, 702–705 (2013).
21. Naef, R. & Seebach, D. Preparation of the enantiomerically pure *cis*-configured and *trans*-configured 2-(*tert*-butyl)-3-methylimidazolidin-4-ones from the amino-acids (S)-alanine, (S)-phenylalanine, (R)-phenylglycine, (S)-methionine, and (S)-valine. *Helv. Chim. Acta* **68**, 135–143 (1985).
22. Holden, C. M. & Greaney, M. F. Modern aspects of the Smiles rearrangement. *Chem. Eur. J.* **23**, 8992–9008 (2017).
23. Snape, T. J. A truce on the Smiles rearrangement: revisiting an old reaction—the Truce–Smiles rearrangement. *Chem. Soc. Rev.* **37**, 2452–2458 (2008).
24. Sung, R.-Y. et al. Kinetic studies on the nucleophilic substitution reaction of 4-X-substituted-2,6-dinitrochlorobenzene with pyridines in MeOH–MeCN mixtures. *Bull. Korean Chem. Soc.* **30**, 1579–1582 (2009).
25. Bunnett, J. F. & Zahler, R. E. Aromatic nucleophilic substitution reactions. *Chem. Rev.* **49**, 273–412 (1951).
26. Schimmler, S. D. et al. Nucleophilic deoxyfluorination of phenols via aryl fluorosulfonate intermediates. *J. Am. Chem. Soc.* **139**, 1452–1455 (2017).
27. Neumann, C. N. & Ritter, T. Facile C–F bond formation through a concerted nucleophilic aromatic substitution mediated by the PhenoFluor reagent. *Acc. Chem. Res.* **50**, 2822–2833 (2017).
28. Kwan, E. E., Zeng, Y., Besser, H. A. & Jacobsen, E. N. Concerted nucleophilic aromatic substitutions. *Nat. Chem.* **10**, 917–923 (2018).
29. Clayden, J., Hennecke, U., Vincent, M. A., Hillier, I. H. & Helliwell, M. The origin of the conformational preference of *N,N'*-diaryl-*N,N'*-dimethyl ureas. *Phys. Chem. Chem. Phys.* **12**, 15056–15064 (2010).
30. Costil, R. et al. Heavily substituted atropisomeric diarylamines by unactivated Smiles rearrangement of *N*-aryl anthranilamides. *Angew. Chem. Int. Ed.* **56**, 12533–12537 (2017).

Acknowledgements We acknowledge funding from the EPSRC (GR/L018527) and ERC (Advanced Grant ROCOCO and Proof of Concept grant QUATERMAIN), and we are grateful to M. M. Amer for assistance with the synthesis of starting materials.

Reviewer information Nature thanks T. Kawabata and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.J.L., J.W.W. and J.C. devised the experiments; D.J.L. and J.W.W. carried out the experiments; D.J.L., J.W.W. and J.C. analysed the results and wrote the paper.

Competing interests The authors have filed a patent on this work (GB1621512.1).

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0553-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0553-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended Data Table 1 | Optimizing the synthesis of 3

entry	starting materials	solvent ^a	base ^b	additive	product	yield
1	<i>cis</i> - 1-Ala + PhNHMe	CH ₂ Cl ₂	Et ₃ N	—	<i>cis</i> - 3-Ala-a	95
2	<i>trans</i> - 1-Ala + PhNHMe	CH ₂ Cl ₂	Et ₃ N	—	—	0
3	<i>trans</i> - 1-Ala + PhNHMe	CH ₂ Cl ₂	Et ₃ N	KI ^c	<i>trans</i> - 3-Ala-a	40
4	<i>trans</i> - 1-Ala + PhNHMe	CH ₂ Cl ₂ ^d	Et ₃ N	KI ^c	<i>trans</i> - 3-Ala-a	86
5	<i>cis</i> - 1-Phe + PhNHMe	CH ₂ Cl ₂	Et ₃ N	KI ^c	<i>cis</i> - 3-Phe-a	90
6	<i>trans</i> - 1-Phe + PhNHMe	CH ₂ Cl ₂	Et ₃ N	KI ^c	<i>trans</i> - 3-Phe-a	20
7	2-Ala + PhMeNCOCi	MeCN	—	—	<i>cis</i> - 3-Ala-a	50
8	2-Ala + PhMeNCOCi	PhMe	—	—	<i>cis</i> - 3-Ala-a	65
9	2-Ala + PhMeNCOCi	PhMe	Et ₃ N	—	<i>cis</i> - 3-Ala-a	<5
10 ^e	2-Ala + PhMeNCOCi	PhMe	—	DMAP ^f	<i>cis</i> - 3-Ala-a	88
11	2-Ala + PhMeNCOCi	(CH ₂ Cl ₂) ₂	—	DMAP ^f	<i>cis</i> - 3-Ala-a	85

^aReaction carried out at reflux for 18 h unless otherwise indicated.^b1.5 equiv.^c1.1 equiv.^d45 h.^eMethod C.^f0.05 equiv.

Extended Data Table 2 | Optimizing the rearrangement of 3 to 4

entry	Starting material	Base ^a	<i>T</i> / °C	product	yield	<i>er</i>
1	<i>trans</i> - 3-Ala-a	LDA	−78	–	0 (96 ^b)	– ^c
2	<i>trans</i> - 3-Ala-a	LDA	0 - rt	4-Ala-a	95	– ^c
3	<i>cis</i> - 3-Ala-a	LDA	0 - rt	<i>ent</i> - 4-Ala-a	92	– ^c
4 ^d	<i>trans</i> - 3-Ala-a	KHMDS	0 - rt	4-Ala-a	95 ^e	>99:1
5 ^d	<i>cis</i> - 3-Ala-a	KHMDS	0 - rt	<i>ent</i> - 4-Ala-a	99	<1:99

^a1.5 equiv.^bYield of *ent-cis*-**3-Ala-a** formed by epimerization.^cNot determined.^dMethod A.^eReaction on 1.5-g scale.

Widespread seasonal compensation effects of spring warming on northern plant productivity

Wolfgang Buermann^{1,2*}, Matthias Forkel³, Michael O'Sullivan¹, Stephen Sitch⁴, Pierre Friedlingstein⁵, Vanessa Haverd⁶, Atul K. Jain⁷, Etsushi Kato⁸, Markus Kautz⁹, Sebastian Lienert^{10,11}, Danica Lombardozzi¹², Julia E. M. S. Nabel¹³, Hanqin Tian^{14,15}, Andrew J. Wiltshire¹⁶, Dan Zhu¹⁷, William K. Smith¹⁸ & Andrew D. Richardson^{19,20}

Climate change is shifting the phenological cycles of plants¹, thereby altering the functioning of ecosystems, which in turn induces feedbacks to the climate system². In northern (north of 30° N) ecosystems, warmer springs lead generally to an earlier onset of the growing season^{3,4} and increased ecosystem productivity early in the season⁵. In situ⁶ and regional^{7–9} studies also provide evidence for lagged effects of spring warmth on plant productivity during the subsequent summer and autumn. However, our current understanding of these lagged effects, including their direction (beneficial or adverse) and geographic distribution, is still very limited. Here we analyse satellite, field-based and modelled data for the period 1982–2011 and show that there are widespread and contrasting lagged productivity responses to spring warmth across northern ecosystems. On the basis of the observational data, we find that roughly 15 per cent of the total study area of about 41 million square kilometres exhibits adverse lagged effects and that roughly 5 per cent of the total study area exhibits beneficial lagged effects. By contrast, current-generation terrestrial carbon-cycle models predict much lower areal fractions of adverse lagged effects (ranging from 1 to 14 per cent) and much higher areal fractions of beneficial lagged effects (ranging from 9 to 54 per cent). We find that elevation and seasonal precipitation patterns largely dictate the geographic pattern and direction of the lagged effects. Inadequate consideration in current models of the effects of the seasonal build-up of water stress on seasonal vegetation growth may therefore be able to explain the differences that we found between our observation-constrained estimates and the model-constrained estimates of lagged effects associated with spring warming. Overall, our results suggest that for many northern ecosystems the benefits of warmer springs on growing-season ecosystem productivity are effectively compensated for by the accumulation of seasonal water deficits, despite the fact that northern ecosystems are thought to be largely temperature- and radiation-limited¹⁰.

Northern land regions have experienced substantial warming since the early 1970s, which has changed how ecosystems function¹¹. One prominent example of emerging ecosystem responses is shifts in plant phenological cycles: earlier spring onset and delayed autumn senescence are lengthening the northern growing season^{6,12}. These phenological shifts have altered ecosystem productivity^{5,6,8,13,14} and the seasonality of important ecosystem feedbacks to the atmosphere and climate system^{6,15}.

Warmer and earlier springs may also influence ecosystem function later in the growing season through indirect or lagged effects^{16,17}.

For example, in situ studies provide evidence for substantial positive lagged effects on ecosystem productivity, whereby the influence of warmer springs may be conveyed to subsequent seasons through the development of larger leaves or increased foliar nitrogen⁶. By contrast, warmer or earlier springs may cause earlier autumn senescence because of the fixed lifespans of leaves¹⁸ or adversely affect plant productivity later in the season through the build-up of water deficits^{7–9,19,20}. However, a more comprehensive understanding of lagged productivity responses is still lacking.

Here, we use long-term (spanning the period 1982–2011) satellite data of vegetation greenness (as a proxy for potential photosynthesis)²¹, flux-tower and model estimates of CO₂ uptake through photosynthesis (gross primary productivity, GPP)^{22,23} and high-resolution climate data²⁴ to estimate the strength and geographic distribution of lagged effects that capture the influence of spring phenological transitions on plant productivity during the subsequent summer and autumn. Our analysis relies on identifying correlations between spring temperature (which serves as an independent phenological indicator) and satellite greenness or simulated GPP during spring and subsequent seasons to estimate concurrent phenological responsiveness and linked lagged effects (see Methods).

Across northern land, correlations between annual spring temperature and spring greenness show a significantly positive and spatially extensive pattern consistent with the notion of a tight control of spring temperature on concurrent plant productivity: 80% of northern (north of 30° N) vegetated non-agricultural land (total study area of roughly 41 × 10⁶ km²) exhibits statistically significant ($P < 0.05$ at the grid-cell level) positive correlations (Fig. 1a). To assess lagged effects on plant productivity associated with anomalous spring temperatures, we computed partial correlations between spring temperature and subsequent summer and autumn greenness, whereby covarying effects of concurrent climate on these correlations are controlled for (see Supplementary Information, section 1). Partial correlations between annual spring temperature and subsequent summer greenness show a widespread positive (6%, $P < 0.05$) and negative (6%, $P < 0.05$) pattern (Fig. 1b). Areas of positive partial correlations are predominantly situated in Eurasia, covering vast regions north of 50° N, whereas areas with negative correlations are more localized in western North America, Siberia and temperate eastern Asia. The partial correlation pattern between spring temperature and autumn greenness indicates an extension of the summer pattern of negative correlation (11%, $P < 0.05$; positive correlations cover only 2%, $P < 0.05$), with additional coverage seen

¹Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds, UK. ²Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA, USA. ³Climate and Environmental Remote Sensing Group, Department for Geodesy and Geoinformation, TU Wien, Vienna, Austria. ⁴College of Life and Environmental Sciences, University of Exeter, Exeter, UK. ⁵College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK. ⁶CSIRO Oceans and Atmosphere, Canberra, Australian Capital Territory, Australia. ⁷Department of Atmospheric Sciences, University of Illinois, Urbana, IL, USA. ⁸Institute of Applied Energy, Tokyo, Japan. ⁹Forest Research Institute Baden-Württemberg, Freiburg, Germany. ¹⁰Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland. ¹¹Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland. ¹²National Center for Atmospheric Research, Climate and Global Dynamics, Terrestrial Sciences Section, Boulder, CO, USA. ¹³Max Planck Institute for Meteorology, Hamburg, Germany. ¹⁴International Center for Climate and Global Change Research, School of Forestry and Wildlife Sciences, Auburn University, Auburn, AL, USA. ¹⁵Research Center for Eco-Environmental Sciences, State Key Laboratory of Urban and Regional Ecology, Chinese Academy of Sciences, Beijing, China. ¹⁶Met Office Hadley Centre, Exeter, UK. ¹⁷Laboratoire des Sciences du Climat et de l'Environnement, LSCE CEA-CNRS-UVSQ, Gif sur Yvette, France. ¹⁸School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, USA. ¹⁹School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA. ²⁰Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA. *e-mail: wolfgang.buermann@geo.uni-augsburg.de

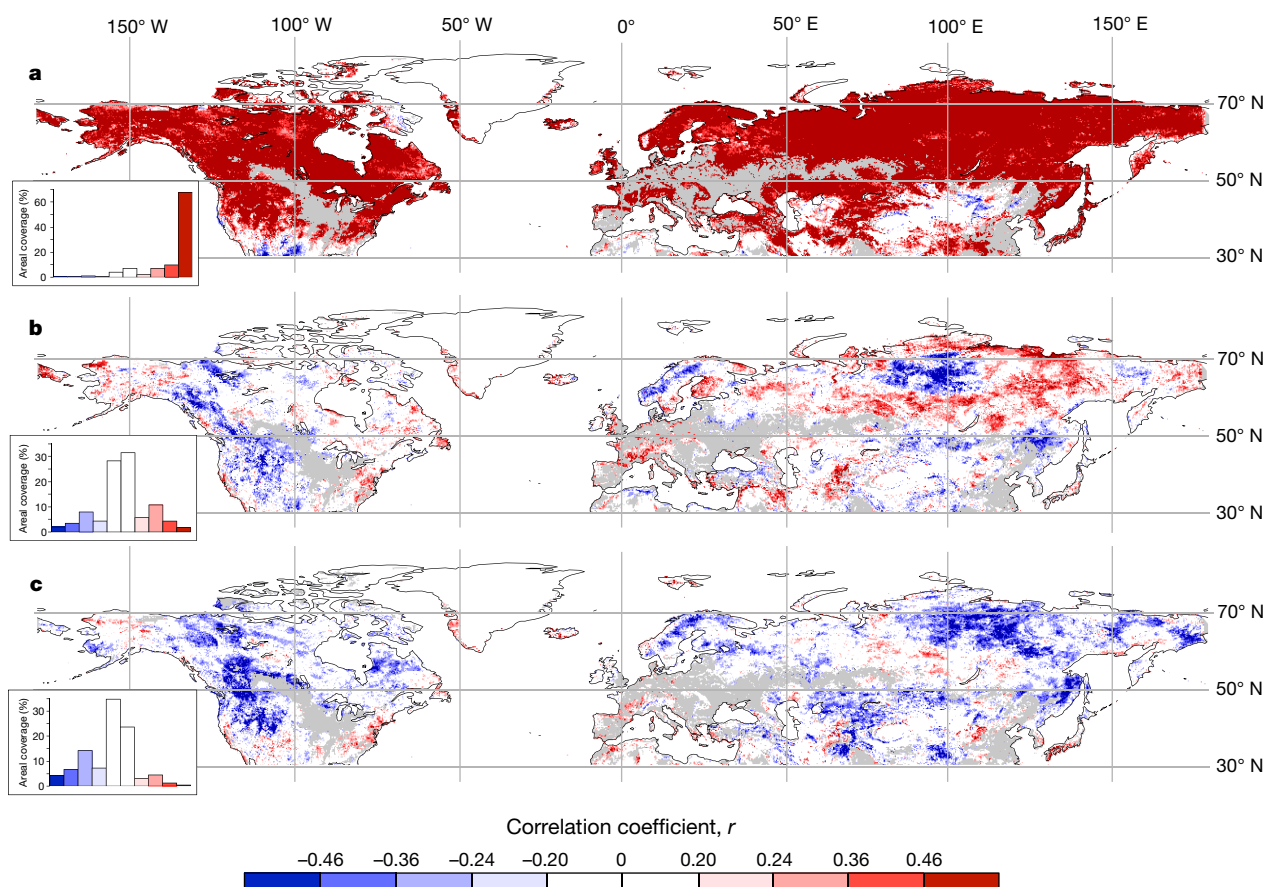


Fig. 1 | Spatial pattern of concurrent and lagged productivity responses to spring warming based on satellite greenness observations. **a**, Grid-cell correlations between yearly spring temperatures and spring satellite vegetation greenness (expressed through the normalized difference vegetation index, NDVI) for our study period, 1982–2011. **b**, **c**, Partial correlations between annual spring temperature and subsequent summer (**b**) and autumn (**c**) NDVI over this period. In these partial correlations, the covarying influences of summer temperature and precipitation (**b**) and autumn temperature and precipitation (**c**) on the correlations between

spring temperature and summer or autumn NDVI have been removed.

Seasons are defined using a local adaptive procedure (see Methods).

Absolute values of the correlation coefficient (r) correspond to significance levels of $P = 0.3$ ($r = 0.20$), $P = 0.2$ ($r = 0.24$), $P = 0.05$ ($r = 0.36$) or $P = 0.01$ ($r = 0.46$). For each map, frequency histograms showing the areal coverage corresponding to positive and negative correlations, estimated as a fraction of total study area, are also provided (see insets). Areas that are cultivated or managed³⁰ (light grey) are not included in the analysis.

mainly in northeastern Eurasia and temperate central Asia (Fig. 1b, c). Although long-term trends in temperature and greenness could potentially influence these correlations, a corresponding analysis on detrended data shows that the patterns are similar (Supplementary Information, section 1). This similarity suggests a dominant influence of interannual to quasi-decadal variability on the correlation pattern between spring temperature and satellite greenness during subsequent seasons. A comparison of the strength of these lagged relationships with concurrent climatic influences on greenness pattern shows that at regional scales the influence of spring temperature on summer and autumn greenness can be equally important or even dominant (Supplementary Information, section 1).

To further assess the robustness of the lagged productivity responses that we identified from satellite data, we compared these responses to those inferred from flux-tower measurements of land–atmosphere CO_2 flux (FLUXNET). The results show that, across $n = 16$ tower sites, the strength and direction of relationships between spring temperature and spring and summer greenness derived from satellite data correspond well to those based on spring temperature and spring and summer GPP derived from tower data (Extended Data Fig. 1). However, the agreement between the relationships between spring temperature and autumn satellite greenness and between spring temperature and autumn tower-derived GPP is not as strong (Extended Data Fig. 1). This validation has several caveats, including the small number of available tower sites and the differences in spatial scales for satellite

(coarse) and tower (fine-scale) data; however, the overall consistency in the estimated lagged productivity responses suggests that the estimates based on satellite data are plausible.

The geographic distribution of the relationship between changes in spring temperature and subsequent summer greenness (see Fig. 1b) suggests that some combination of climate, elevation and land cover may explain these patterns. To investigate this, we conducted a random-forest analysis using a set of predictors that encapsulate such factors (see Supplementary Information, section 2). The results show that the partial correlation pattern between spring temperature and summer greenness can be explained with elevation and selected climate variables (such as summer precipitation and precipitation seasonality) acting as the most important variables (Extended Data Fig. 2, Supplementary Information, section 2). Across northern ecosystems, we find that these partial correlations tend to become more negative with higher elevation, but such well-defined directional relationships are not as apparent for important precipitation metrics (Extended Data Fig. 2).

Grouping the lagged productivity responses on the basis of the direction of robust correlations between spring temperature and spring, summer and autumn greenness reveals large clusters of regions with negative lagged effects and more scattered areas with positive lagged effects (Fig. 2a). As a result, negative lagged productivity responses associated with spring warming and greening, coupled with declines in summer or autumn greenness, stretch over vast areas in western North America, Siberia and to some extent eastern temperate Asia, whereas

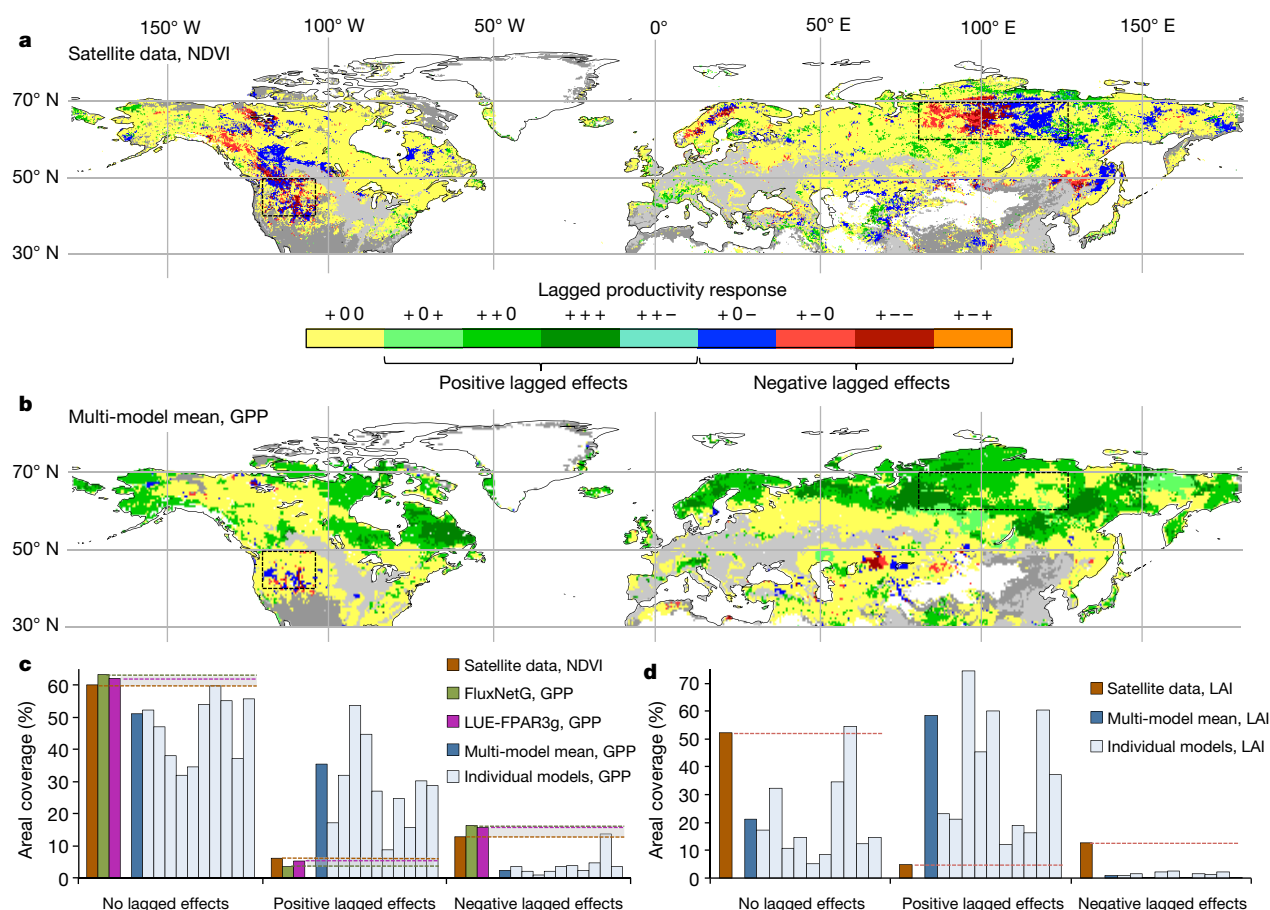


Fig. 2 | Spatial pattern of lagged productivity responses based on satellite greenness observations and models. **a, b,** The maps summarize the direction of robust ($P < 0.05$) grid-cell correlations between annual spring temperature and spring, summer and autumn NDVI determined from satellite data (**a**) or spring, summer and autumn GPP determined from the TRENDYv6 multi-model mean (**b**). For example, the lagged productivity response denoted as '+-0' represents positive correlations between spring temperature and spring NDVI or GPP, negative correlations between spring temperature and summer NDVI or GPP, and no correlations between spring temperature and autumn NDVI or GPP. The relationships between spring temperature and summer as well as autumn NDVI or GPP are estimated using partial correlations, whereby effects of covarying concurrent climate influences have been controlled for (see Fig. 1, Methods). The corresponding patterns for individual models are shown in Extended Data Fig. 3. Areas with no robust link between spring temperature and spring NDVI or GPP (dark grey) and areas that are cultivated or managed (light grey) are also shown. The two focal

regions in this study (western USA and Siberia) are indicated by black-dashed rectangles. **c,** Extent of areas with no, positive or negative lagged effects (see definition in **a**) within the study region for satellite NDVI data (brown) and GPP based on TRENDYv6 models (dark blue, multi-model mean; light blue, individual models). Corresponding results from a similar analysis for two satellite-data-constrained GPP datasets, based on upscaled FLUXNET data (FluxNetG; green) and a light-use-efficiency model (LUE-FPAR3g; magenta; see Methods), are also shown (see also Extended Data Fig. 4). The horizontal dashed lines are to aid comparison to the TRENDYv6 model results and the shaded regions encapsulate the spread among the three estimates derived from satellite-based approaches. **d,** Results from a complementary analysis for satellite-data-constrained and modelled LAI (see Methods). Results from the same analysis for detrended data show that the differences between observation- and model-based estimates of the areal fractions of positive and negative lagged effects are similar (Supplementary Information, section 1).

positive lagged effects are more common in eastern Eurasia north of 50° N (except Siberia).

Carbon-cycle models must be able to simulate the responses of vegetation phenology and the corresponding effects on ecosystem productivity and net carbon uptake realistically to estimate climate-carbon feedbacks credibly²⁵. We therefore assessed the ability of ten current-generation models that contribute to TRENDYv6^{22,23} to replicate the observed lagged productivity responses to spring warming. The results reveal a substantially higher multi-model mean areal coverage of positive lagged effects on plant productivity (and much lower coverage of negative lagged effects) than for the satellite estimates (Fig. 2a, b). Although there are marked differences among the individual models (Extended Data Fig. 3), a notable pattern in the ensemble is the near absence of any negative lagged effects across Siberia and the overall abundance of positive lagged effects that extend over summer and autumn (Fig. 2a, b). Satellite greenness has been used extensively as a proxy for vegetation productivity^{3,26}, but direct comparisons between greenness and GPP patterns are limited (see Methods). However,

a similar analysis using two satellite-data-constrained GPP datasets (based on upscaled FLUXNET data and a light-use-efficiency model; see Methods) reveals nearly identical lagged productivity patterns to those based on satellite greenness (Extended Data Fig. 4).

Grouping the lagged productivity responses more broadly, into positive and negative lagged effects, yields an areal extent of regions with positive lags of 36% for the TRENDYv6 ensemble (9%–54% for the ten individual TRENDYv6 models) and 4%–6% for the estimates derived from satellite data and satellite-data-constrained approaches (Fig. 2c). The areal coverage of negative lagged effects predicted by the TRENDYv6 ensemble is only 2% (1%–14% for the ten models), whereas that estimated from the satellite-based approaches is 13%–16%. (The ranges for the satellite-based estimates encapsulate the spread among the three different estimates; see shading in Fig. 2c.)

It is not clear why these terrestrial carbon-cycle models cannot adequately replicate the observed spatial pattern of lagged productivity responses to warmer springs. One key factor could be how seasonal vegetation growth is represented in the models. To assess this,

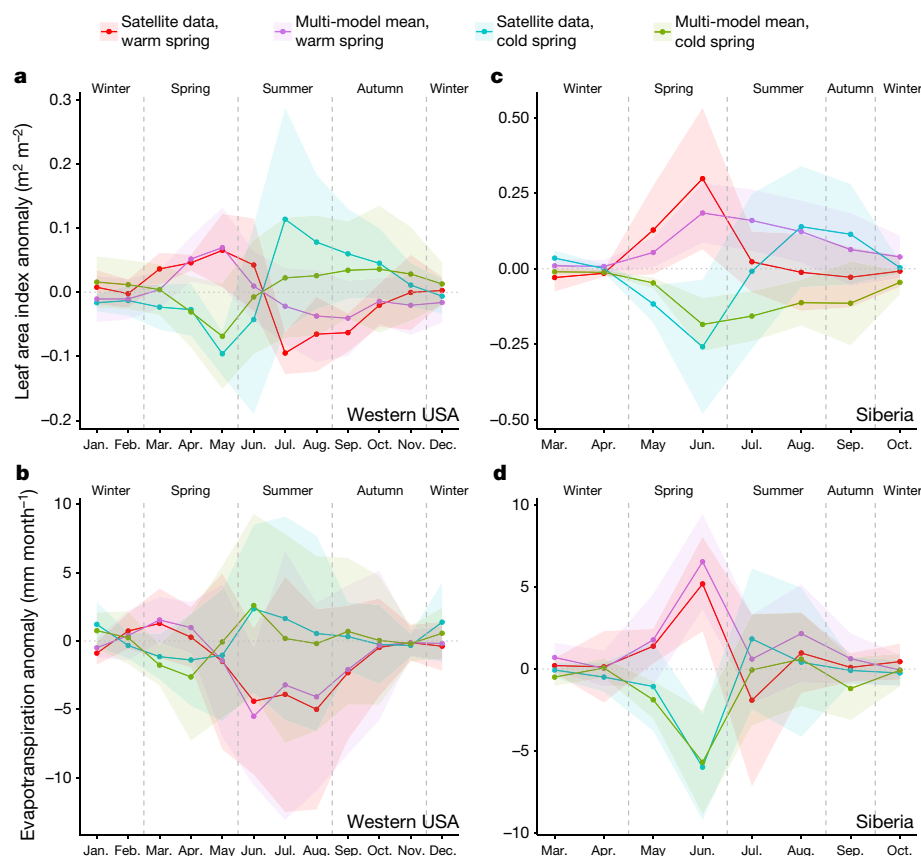


Fig. 3 | Seasonal trajectories of regionally averaged LAI and evapotranspiration anomalies based on observation-constrained and modelling approaches for warm- and cold-spring years. **a–d**, Monthly anomalies in spatially averaged and maximum composited LAI (**a**, **c**) and evapotranspiration (**b**, **d**) based on satellite-data-constrained estimates (LAI3g, ET-GLEAM) and model simulations (TRENDYv6 multi-model mean) for western USA (**a**, **b**) and Siberia (**c**, **d**). Western USA encompasses the non-agricultural regions from 120° W to 105° W and 40° N to 50° N, whereas Siberia is defined to be

from 80° E to 125° E and 60° N to 70° N (see also Fig. 2). Anomalies are relative to the mean over the study period, 1982–2011. The monthly maximum composites shown are based on the mean LAI or evapotranspiration of the seven warmest- and coldest-spring years within the study period. The climatological seasons are indicated by the vertical grey dashed lines. Uncertainty bounds (shaded areas) reflect the spread in the monthly LAI or evapotranspiration anomalies within the compositing period (± 1 s.d., $n = 7$).

we performed a similar seasonal correlation analysis with satellite and modelled leaf area index (LAI) data (see Methods). The results reveal an even larger discrepancy between the areal proportions of positive and negative lagged responses of LAI to spring warming determined using observation-based and modelling approaches compared to those determined using productivity metrics (Fig. 2c, d, Extended Data Fig. 4). The substantial overestimation of growing-season LAI in the models in response to spring warmth could cause too much new carbon to be allocated in plant tissue, which then enhances GPP.

Limited water availability may cause adverse lagged effects in response to spring warmth and could help to reconcile the differences between observations and models. To further investigate this we performed a regional analysis for western USA and Siberia, for which observation-based and simulated lagged productivity responses show more converging and diverging patterns, respectively (see Fig. 2). For western USA, we find that seasonal trajectories in aggregated satellite-data-constrained and modelled LAI and evapotranspiration display positive anomalies during spring in years with warmer springs and corresponding negative anomalies later in the growing season (suggestive of negative lagged effects associated with a build-up of water stress) (Fig. 3a, b). However, for Siberia, the seasonal trajectories in observation-constrained and modelled LAI for warm-spring years start to diverge substantially during summer and autumn, with the observations displaying more negative anomalies during summer and autumn (again suggestive of water stress) and the opposite pattern predicted by the models (Fig. 3c). Seasonal trajectories of observation-based and modelled evapotranspiration for years with anomalous

spring temperatures are more in agreement, although there is some indication that the models underestimate water stress in summer in warm-spring years (Fig. 3d). The consistency between the observed and modelled responses of LAI and evapotranspiration to spring warmth over western USA, a region that is known for its vulnerability to drought in response to spring warmth^{27–29}, suggests that the hydrology and phenology schemes included in the models are generally fit for purpose. The strong divergence between observation-based and modelled responses of seasonal vegetation growth to spring warmth over Siberia (which is dominated by needleleaf deciduous forests) may be due to underestimation of the effects of water stress on seasonal canopy development and general omission of fixed leaf lifespans in the models (Extended Data Table 1, Supplementary Information, section 3). We estimate that, owing to the difference between observation-based and modelled productivity responses to anomalous spring temperatures across Siberia, annual GPP for a warm-spring year may be up to four times higher in the TRENDYv6 ensemble (1.7 Pg C yr^{-1}) than an observation-constrained estimate based on upscaled FLUXNET data (0.4 Pg C yr^{-1}) (Extended Data Fig. 5).

Our analysis based on satellite vegetation records over multiple decades provides evidence for widespread positive and negative lagged plant-productivity responses across northern ecosystems associated with warmer springs. The spatially extensive pattern of negative lagged effects that we identified implies substantially reduced benefits for ecosystem productivity and carbon sequestration from longer northern growing seasons under climate change. We have also shown that current terrestrial carbon-cycle models substantially underestimate

(overestimate) negative (positive) lagged effects associated with spring warming. This is possibly because these models inadequately capture the effects of the seasonal build-up of water stress on seasonal vegetation growth. Continued monitoring of emerging ecosystem responses and improved modelling capabilities will therefore be crucial to improve our understanding of the complex interactions between a changing climate, shifts in phenological cycles and effects on energy, water and carbon cycles.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0555-7>.

Received: 9 March 2018; Accepted: 8 August 2018;

Published online 3 October 2018.

1. Fu, Y. H. et al. Declining global warming effects on the phenology of spring leaf unfolding. *Nature* **526**, 104–107 (2015).
2. Peñuelas, J., Rutishauser, T. & Filella, I. Phenology feedbacks on climate change. *Science* **324**, 887–888 (2009).
3. Myneni, R., Keeling, C., Tucker, C., Asrar, G. & Nemani, R. R. Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature* **386**, 698–702 (1997).
4. Menzel, A. et al. European phenological response to climate change matches the warming pattern. *Glob. Change Biol.* **12**, 1969–1976 (2006).
5. Keenan, T. F. et al. Net carbon uptake has increased through warming-induced changes in temperate forest phenology. *Nat. Clim. Change* **4**, 598–604 (2014).
6. Richardson, A. D. et al. Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agric. For. Meteorol.* **169**, 156–173 (2013).
7. Grippa, M. et al. The impact of snow depth and snowmelt on the vegetation variability over central Siberia. *Geophys. Res. Lett.* **32**, L21412 (2005).
8. Buermann, W., Bikash, P. R., Jung, M., Burn, D. H. & Reichstein, M. Earlier springs decrease peak summer productivity in North American boreal forests. *Environ. Res. Lett.* **8**, 024027 (2013).
9. Sippel, S. et al. Contrasting and interacting changes in simulated spring and summer carbon cycle extremes in European ecosystems. *Environ. Res. Lett.* **12**, 075006 (2017).
10. Nemani, R. R. et al. Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *Science* **300**, 1560–1563 (2003).
11. Settele, J. et al. in *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the IPCC* (eds Field, C. B. et al.) 271–359 (Cambridge Univ. Press, Cambridge, 2014).
12. Forkel, M. et al. Codominant water control on global interannual variability and trends in land surface phenology and greenness. *Glob. Change Biol.* **21**, 3414–3435 (2015).
13. Piao, S. L. et al. Net carbon dioxide losses of northern ecosystems in response to autumn warming. *Nature* **451**, 49–52 (2008).
14. Forkel, M. et al. Enhanced seasonal CO₂ exchange caused by amplified plant productivity in northern ecosystems. *Science* **351**, 696–699 (2016).
15. Forzieri, G. et al. Satellites reveal contrasting responses of regional climate to the widespread greening of Earth. *Science* **356**, 1180–1184 (2017).
16. Richardson, A. D. et al. Influence of spring and autumn phenological transitions on forest ecosystem productivity. *Philos. Trans. R. Soc. Lond. B* **365**, 3227–3246 (2010).
17. Wolf, S. et al. Warm spring reduced carbon cycle impact of the 2012 US summer drought. *Proc. Natl Acad. Sci. USA* **113**, 5880–5885 (2016).
18. Keenan, T. & Richardson, A. D. The timing of autumn senescence is affected by the timing of spring phenology: implications for predictive models. *Glob. Change Biol.* **21**, 2634–2641 (2015).
19. Barnett, T. P., Adam, J. C. & Lettenmaier, D. P. Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature* **438**, 303–309 (2005).
20. Zhang, K., Kimball, J. S., Kim, Y. & McDonald, K. C. Changing freeze–thaw seasons in northern high latitudes and associated influences on evapotranspiration. *Hydrol. Processes* **25**, 4142–4151 (2011).
21. Pinzon, J. E. & Tucker, C. J. A. Non-stationary 1981–2012 AVHRR NDVI3g time series. *Remote Sens.* **6**, 6929–6960 (2014).
22. Sitch, S. et al. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences* **12**, 653–679 (2015).
23. LeQuéré, C. et al. Global carbon budget 2017. *Earth Syst. Sci. Data* **10**, 405–448 (2018).
24. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climate observations – the CRU TS3.10 dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
25. Cadule, P. et al. Benchmarking coupled climate-carbon models against long-term atmospheric CO₂ measurements. *Glob. Biogeochem. Cycles* **24**, GB2016 (2010).
26. Huang, M. et al. Velocity of change in vegetation productivity over northern high latitudes. *Nat. Ecol. Evol.* **1**, 1649–1654 (2017).
27. Westerling, A. L., Hidalgo, H. G., Cayan, D. R. & Swetnam, T. W. Warming and earlier spring increase western U.S. forest wildfire activity. *Science* **313**, 940–943 (2006).
28. Sacks, W. J., Schimel, D. S. & Monson, R. K. Coupling between carbon cycling and climate in a high-elevation subalpine forest: a model-data fusion analysis. *Oecologia* **151**, 54–68 (2007).
29. Parida, B. R. & Buermann, W. Increasing summer drying in North American ecosystems in response to longer nonfrozen periods. *Geophys. Res. Lett.* **41**, 5476–5483 (2014).
30. Bartholomé, E. & Belward, A. S. GLC2000: a new approach to global land cover mapping from Earth observation data. *Int. J. Remote Sens.* **26**, 1959–1977 (2005).

Acknowledgements M.O. is funded through an EU Marie Curie Integration grant to W.B. M.F. is funded through the TU Wien Wissenschaftspreis 2015, a personal science award to W. Dorigo. V.H.'s contribution is supported through funding from the Earth Systems and Climate Change Hub of the Australian Government's National Environmental Science Program. H.T. is supported by the National Key R&D Program of China (2017YFA0604702) and the US National Science Foundation (NSF; 1210360, 1243232). A.D.R. is funded through the Macrosystems Biology Program of the NSF (EF-1702697). This work used eddy covariance data acquired and shared by the FLUXNET community, including the following networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia and USCCC. The ERA-Interim reanalysis data were provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization were carried out by the European Fluxes Database Cluster, AmeriFlux Management Project and Fluxdata project of FLUXNET, with the support of the CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. We thank M. Jung for providing upscaled FLUXNET GPP data.

Reviewer information Nature thanks N. Parazoo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.B., M.F. and A.D.R. designed the research. W.B., M.F. and M.O. carried out the analysis and W.B. wrote the manuscript with contributions from all authors. S.S., P.F., V.H., A.K.J., E.K., M.K., S.L., D.L., J.E.M.S.N., H.T., A.J.W. and D.Z. contributed to the TRENDY results. W.K.S. contributed to the LUE-FPAR3g results.

Competing interests : The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0555-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0555-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to W.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data sources. For the satellite vegetation data, we used the GIMMS-NDVI version 3g (NDVI3g)²¹ and LAI3g³¹ products, which are both available at 8-km spatial and 15-day temporal resolution for our study period, 1982–2011. The NDVI3g data stem from optical surface reflectance measurements from a series of NOAA-AVHRR satellites. Effects of orbital drifts, inter-sensor calibration and stratospheric aerosols from volcanic eruption have been corrected for, making this the most consistent long-term satellite vegetation dataset currently available²¹. The LAI3g fields are derived from the NDVI3g data using an artificial neural network model³¹. Gridded monthly climate data were obtained from the Climatic Research Unit (CRU TS3.23) at 0.5° spatial resolution²⁴ for our study period (1982–2011). As an estimate for the observation-constrained evapotranspiration (ET), we included the Global Land Evaporation Amsterdam Model (GLEAM) dataset, which has a spatial resolution of 0.25° at daily time steps³². While the GLEAM approach is based on an empirical model, it is heavily constrained by observations through assimilating satellite microwave vegetation optical depth data as a proxy for water stress³². In addition, land-cover data used in this study are based on the GLC2000 land-cover classification³⁰. For complementary analyses, we also used site-level GPP data derived from the global FLUXNET tower network (FLUXNET2015, tier 1) and two observation-constrained, gridded monthly GPP datasets. The first includes GPP data (0.5° spatial resolution, available for 1982–2008) estimated from upscaled carbon observations based on FLUXNET (FluxNetG)³³. FluxNetG is different from the previously published FluxNet-MTE³³ because it has been produced with inputs from only a single satellite vegetation dataset (NDVIg; a predecessor of NDVI3g) to reduce artefacts from using multiple satellite data (the FluxNetG dataset was also used in ref. ⁸). Second, we used GPP data (0.5° spatial resolution, available for 1982–2011) derived using the light-use-efficiency (LUE) MODIS GPP algorithm driven by bi-monthly GIMMS FPAR3g (LUE-FPAR3g)³⁴. Additional meteorological driver data required as input into the MODIS GPP algorithm were derived from NCEP-DOE Reanalysis II (<http://www.esrl.noaa.gov>). For more information on the GIMMS3g GPP dataset, see ref. ³⁴.

TRENDYv6 models. We also analysed monthly GPP, LAI and ET simulation outputs for 1982–2011 from ten terrestrial carbon-cycle models that were part of a recent model intercomparison project, TRENDYv6^{22,23}. The models included in the analysis here are LPX-Bern, LPJ-GUESS, ISAM, CABLE, VISIT, CLM4.5, DLEM, JSBACH, ORCHIDEE-MICT and JULES. In TRENDYv6, the models were forced with the CRUNCEPv6 climate dataset, which is based on a merged product of the monthly CRU climate data, and to be consistent with the TRENDYv6 ensemble we also used this climate dataset here. In addition, a set of factorial simulations²² were performed and we analysed outputs from a simulation in which only atmospheric CO₂ and climate were varied (land-use change held fixed; experiment 'S2') because our study focus was on non-agricultural ecosystems. For an overview of the processes included in the models relevant to this study, see Extended Data Table 1. For a more general overview of the models see tables 4a and 5 in ref. ²³.

Analysis framework. The satellite bi-monthly GIMMS NDVI3g and LAI3g vegetation data were averaged to a monthly temporal resolution (to be consistent with the TRENDYv6 model outputs). Then, the fine-scale satellite vegetation and coarse-scale CRU temperature fields were (dis)aggregated to a common 0.25° spatial grid on which all correlation analyses were performed. The motivation for this spatial aggregation step is twofold: (i) it retains a certain level of spatial information inherent in the satellite products and (ii) it aligns more closely with the coarser spatial resolutions of the TRENDY carbon-cycle models. Model outputs from TRENDYv6 were either analysed at their native model resolutions spanning grid-cell dimensions from 0.5° to 1.9° (ref. ²²) or resampled to a common 0.5° grid through nearest neighbours (for example, to estimate multi-model means of GPP, LAI and ET at grid-cell levels).

To estimate lagged vegetation growth and productivity responses we first divided the mean seasonal cycle of NDVI or simulated GPP (based on the 30-year study period) into spring, summer and autumn periods for each grid cell. Hereby, the start of spring and the end of autumn are defined by the months in which corresponding temperatures are closest to 0°C, whereas the start and end of summer are defined by the months in which the NDVI (GPP) is closest to 95% (85%) of the annual maximum NDVI (GPP). Alternative approaches for characterizing

phenological cycles involving start and end dates of the growing season are more ambiguous if based solely on optical vegetation indices^{35,36} or when the underlying data have relatively low temporal resolution, as in this study¹².

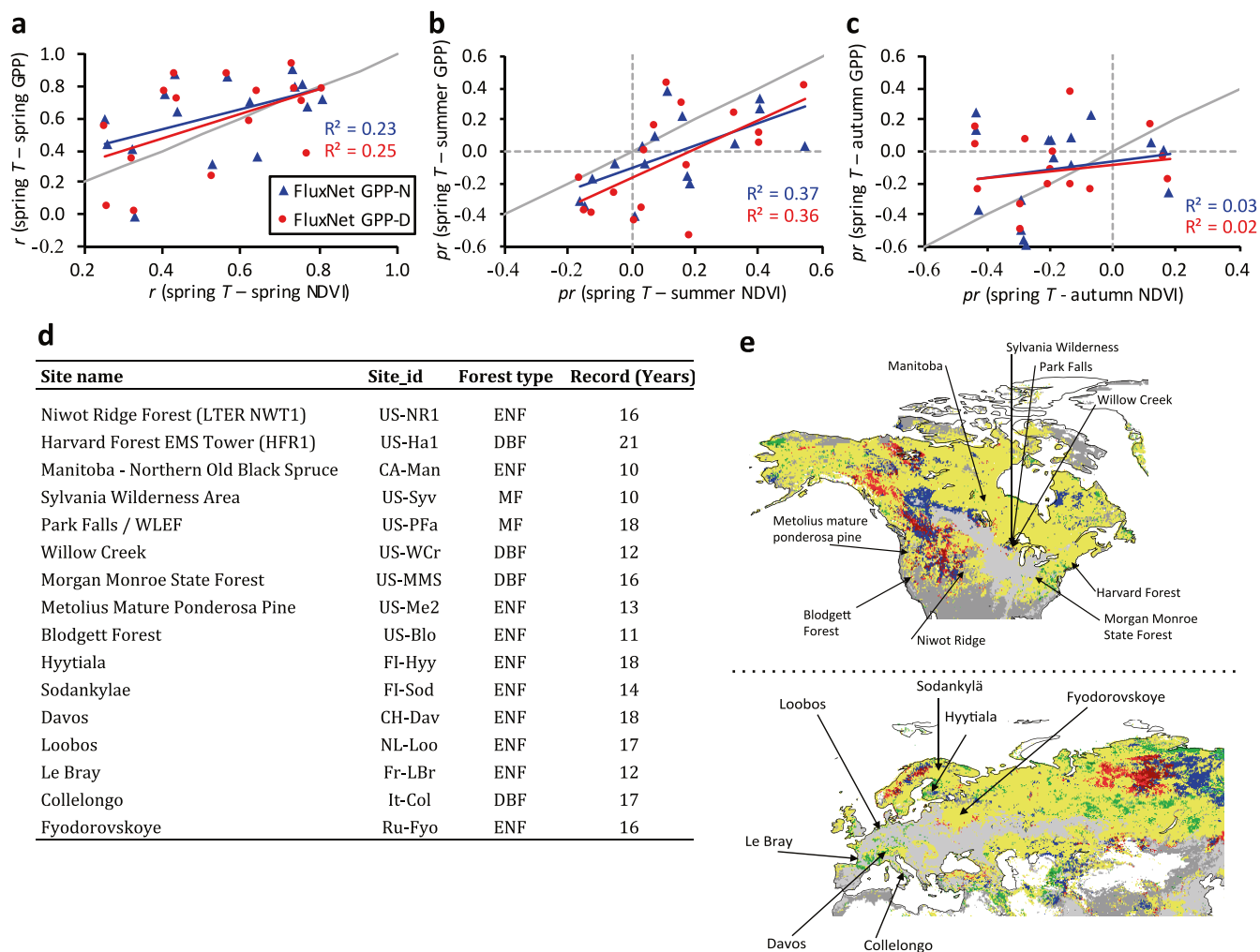
In a next step, we (building on the conceptual model of ref. ¹⁶) classified lagged productivity responses for each grid cell as follows. First, as a minimum requirement for phenological responsiveness to spring warming, we require the spring temperature and the response variable of interest (NDVI, LAI or GPP) to be significantly ($P < 0.05$) positively correlated. Second, we define a lagged productivity (NDVI, GPP) or phenology (LAI) response on the basis of the direction of robust ($P < 0.05$) partial correlations between annual spring temperature (as an independent phenological indicator) and subsequent summer and autumn seasonal means of the response variable of interest; for example, if at a given locality the annual spring temperature is positively correlated with spring NDVI but negatively correlated with subsequent summer NDVI and not robustly correlated with autumn NDVI, then the response label would be '+−0', with the type of symbol denoting the direction of correlations and sequence corresponding to spring–spring, spring–summer and spring–autumn relationships (see Fig. 2). Partial correlations are used to control for covarying effects of climate over seasonal timescales, which can confound the correlations between annual spring temperature and subsequent summer and autumn response variables (see Supplementary Information, section 1).

As indicated, the satellite vegetation data (NDVI3g, LAI3g) used here stem from a series of satellites; although this record has been assembled carefully and validated to some extent³¹, remaining non-vegetation artefacts in the data cannot be ruled out³⁷. Further, satellite greenness (or NDVI) captures the amount of light absorbed by chlorophyll in green leaves³⁸ and has been used extensively as a proxy for spatially resolved vegetation productivity at continental and multi-decadal scales^{3,26}. However, to overcome the limited comparability of directly observed NDVI-based and simulated GPP-based patterns, we also analysed observation-constrained GPP data. The results show good agreement between the lagged productivity patterns (using gridded GPP data from up-scaled FLUXNET and a LUE model), providing further support for the robustness of our results (see Extended Data Fig. 4). Finally, we also used satellite and modelled LAI data to probe the mismatch between lagged greenness and modelled (TRENDYv6) GPP responses to spring warmth.

Data availability

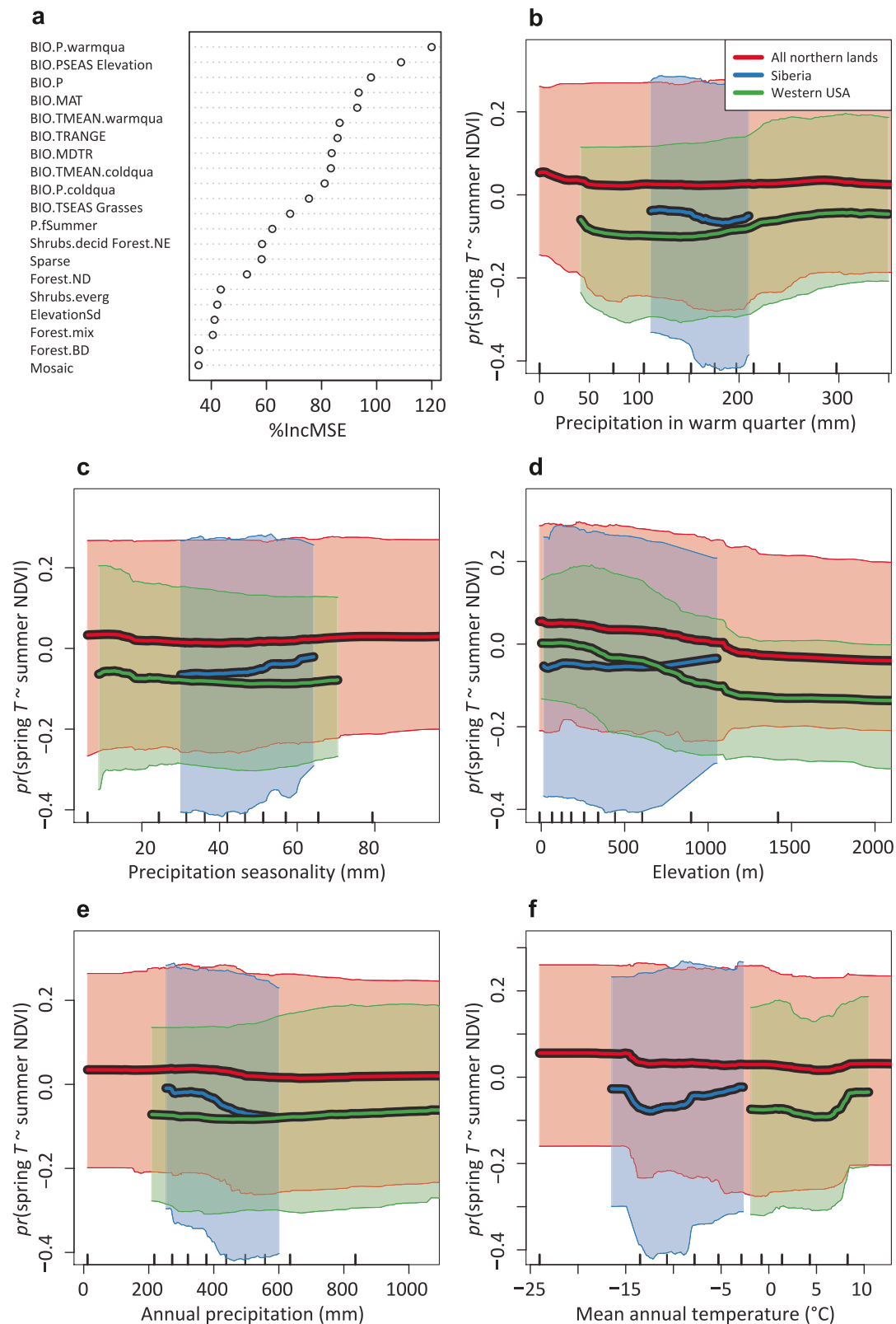
The satellite NDVI3g data that support the findings of this study were downloaded from <http://ecocast.arc.nasa.gov/data/pub/gimms/3g.v0/>. The satellite LAI3g data are available from R. B. Myneni (rmyneni@bu.edu) on reasonable request. The LUE-FPAR3g GPP data can be requested from W.K.S. (wksmith@email.arizona.edu) and the FluxNetG GPP data from M. Jung (mjung@bgc-jena.mpg.de). The TRENDYv6 data are available from S.S. (s.a.sitch@exeter.ac.uk) on reasonable request.

- Zhu, Z. et al. Global data sets of vegetation leaf area index (LAI) 3g and fraction of photosynthetically active radiation (FPAR) 3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3g) for the period 1981 to 2011. *Remote Sens.* **5**, 927–948 (2013).
- Martens, B. et al. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.* **10**, 1903–1925 (2017).
- Jung, M. et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res. Biogeosci.* **116**, G00J07 (2011).
- Smith, W. K. et al. Large divergence of satellite and Earth system model estimates of global terrestrial CO₂ fertilization. *Nat. Clim. Change* **6**, 306–310 (2016).
- Walther, S. et al. Satellite chlorophyll fluorescence measurements reveal large-scale decoupling of photosynthesis and greenness dynamics in boreal evergreen forests. *Glob. Change Biol.* **22**, 2979–2996 (2016).
- Wu, C. et al. Land surface phenology derived from normalized difference vegetation index (NDVI) at global FLUXNET sites. *Agric. For. Meteorol.* **233**, 171–182 (2017).
- Jiang, C. et al. Inconsistencies of interannual variability and trends in long-term satellite leaf area index products. *Glob. Change Biol.* **23**, 4133–4146 (2017).
- Myneni, R. B. et al. The interpretation of spectral vegetation indexes. *IEEE Trans. Geosci. Remote Sens.* **33**, 481–486 (1995).



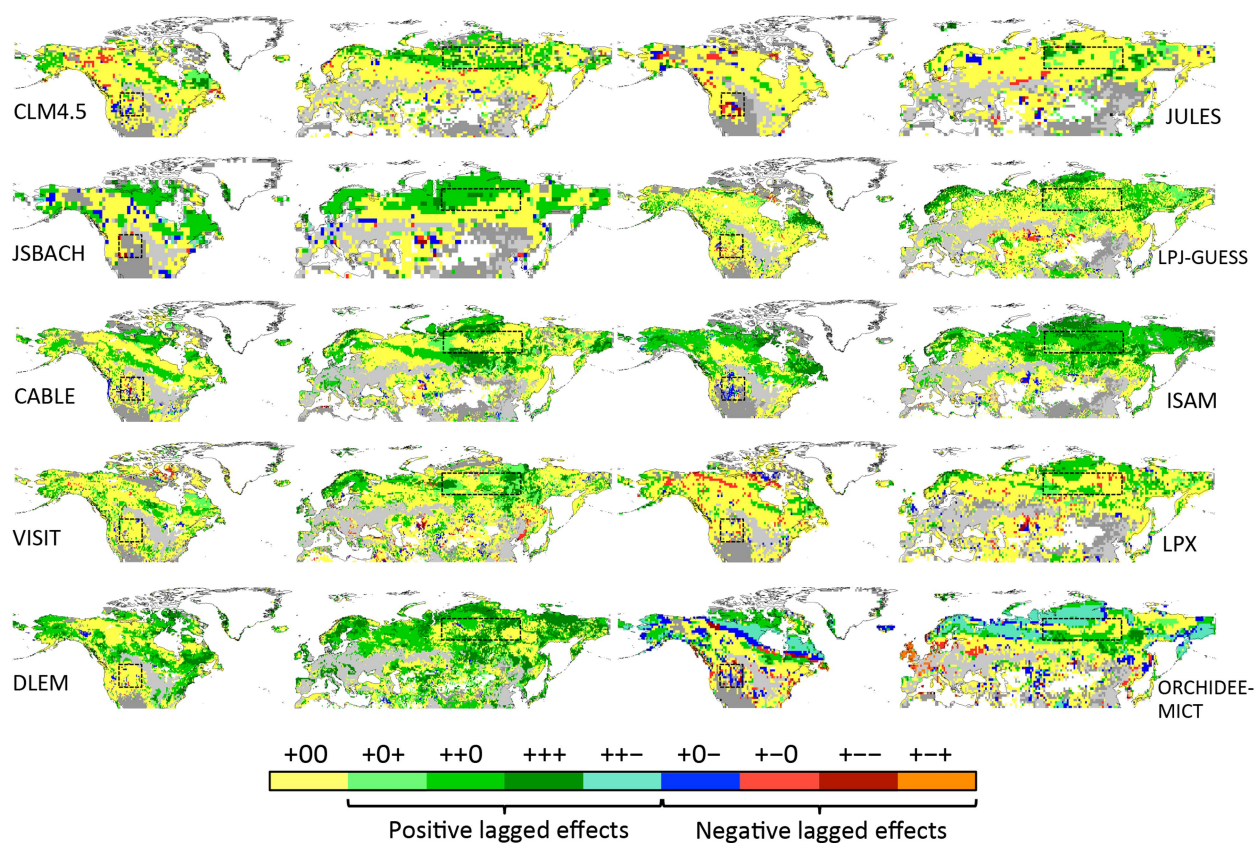
Extended Data Fig. 1 | Comparison of lagged productivity responses based on satellite greenness observations and in situ estimates of carbon fluxes across selected FLUXNET sites. a–c, Site-specific correlations between spring temperature (T) and spring (a), summer (b) or autumn (c) satellite NDVI (x axis) plotted over the corresponding site-specific correlations between spring temperature and spring (a), summer (b) or autumn (c) flux-tower GPP (y axis). In b and c, the relationships are based on partial correlations (pr) between spring temperature and subsequent summer (b) or autumn (c) NDVI or GPP, with covarying effects of summer temperature and precipitation (b) and autumn

temperature and precipitation (c) removed. (Partial) correlations are shown for two estimates of GPP: GPP-N (based on night-time partitioning of net ecosystem exchange) and GPP-D (daytime partitioning). d, For this comparison, satellite NDVI time series at 8-km (native) spatial resolution have been extracted for the 16 FluxNet tower sites with at least 10-year data records. Forest types for the tower sites are: ENF, evergreen needleleaf forest; DBF, deciduous broadleaf forest; MF, mixed forest. e, Maps showing the approximate locations of the FLUXNET tower sites. FLUXNET data for this comparative analysis are from the FLUXNET2015 dataset (tier 1).



Extended Data Fig. 2 | Random-forest analysis to explain the partial correlation pattern between annual spring temperature and summer satellite greenness on hemispheric and regional scales. a, Ranked importance of a set of explanatory variables in a random-forest model for the whole northern ecosystem study region, encompassing all vegetated non-agricultural land north of 30°N (see Supplementary Information, section 2, for details on the explanatory variables used). The ranking is based on the highest increment in mean squared error (IncMSE) between the observed and random-forest-predicted correlation after permuting

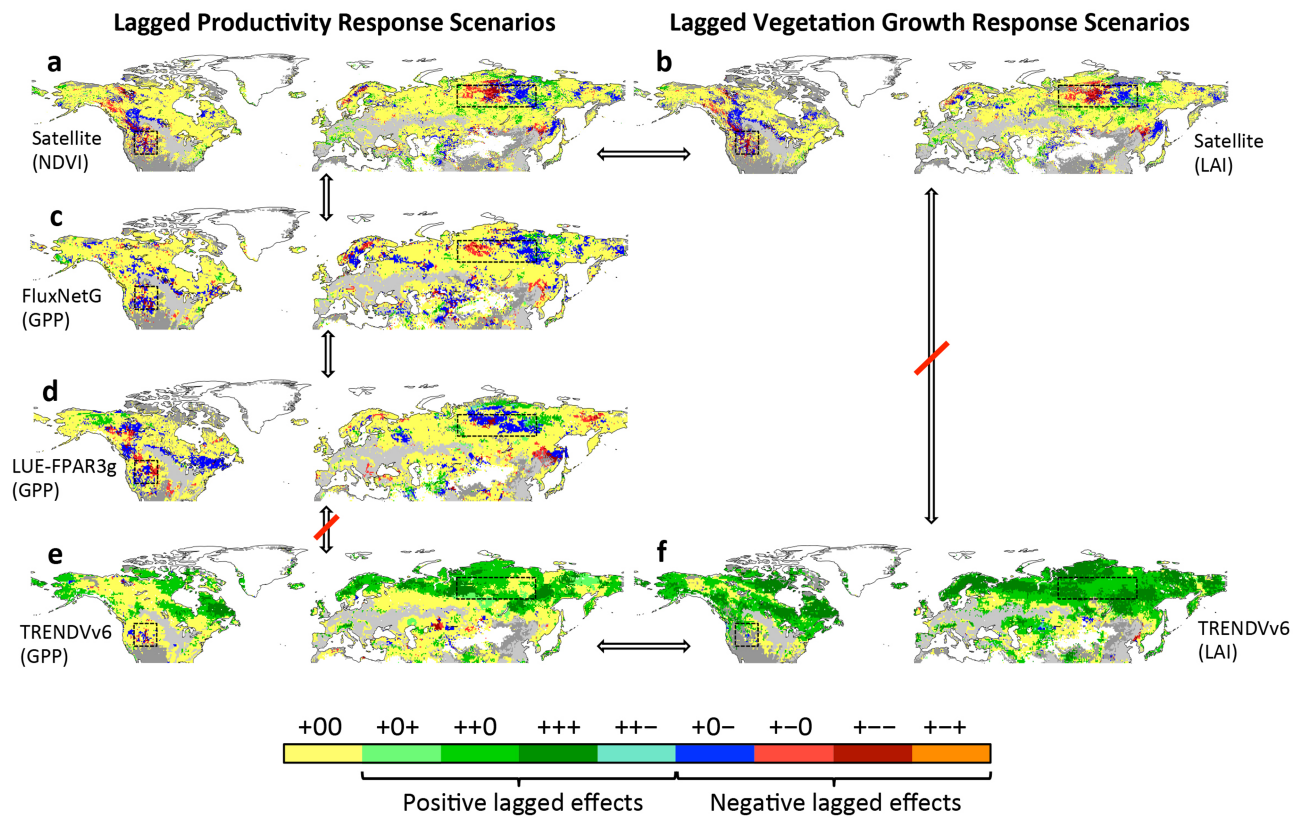
the relevant explanatory variable. **b–f,** Individual conditional expectation lines of the random-forest-predicted partial correlation (pr) between spring temperature (T) and summer NDVI for the five most important explanatory variables. Lines and shaded bands reflect the mean (regional-average response) and the 5%–95% percentile range (grid-cell-level responses to environmental predictors) for the northern (north of 30°N , non-managed) study region (red) and for the focus regions (Siberia, blue; western USA, green) (see Supplementary Information, section 2).



Extended Data Fig. 3 | Spatial pattern of lagged productivity responses based on the individual carbon-cycle models included in TRENDYv6.

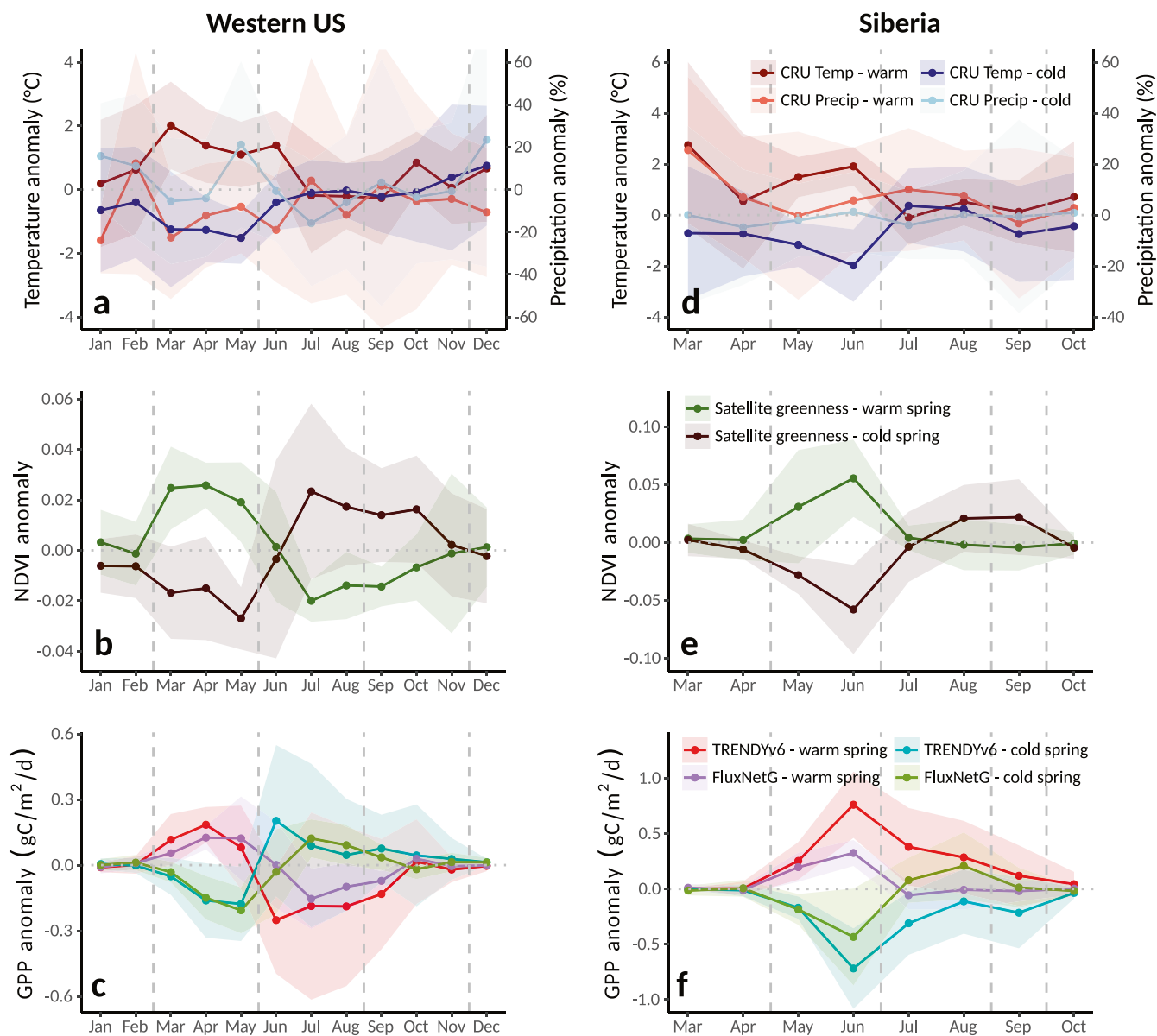
All patterns are based on monthly GPP over the period 1982–2011, using outputs from the ten TRENDYv6 models included in the analysis (see Methods). The maps summarize the direction of statistically

significant ($P < 0.05$) correlation between annual spring temperature and spring, summer or autumn GPP. For details on classification scenarios and contour labels, see Fig. 2. Areas with no robust link between spring temperature and spring GPP (dark grey) and areas that are cultivated or managed (light grey) are also shown.



Extended Data Fig. 4 | Spatial pattern of lagged productivity and vegetation growth responses estimated through satellite-based and modelling approaches. **a–f**, Summary of the direction of robust ($P < 0.05$) correlations between annual spring temperature and spring, summer or autumn satellite NDVI (**a**), satellite LAI (**b**), satellite upscaled GPP (FluxNetG; **c**), satellite-data-driven LUE-modelled GPP (LUE-FPAR3g; **d**),

and multi-model mean GPP (**e**) and LAI (**f**) based on the ten TRENDYv6 models. For details on scenario classifications and contour labels see Fig. 2. Arrows (arrows with strikethroughs) linking panels highlight qualitative agreement (disagreement) between the lagged responses of productivity and vegetation growth based on the various approaches.



Extended Data Fig. 5 | Changes in regional climate, satellite greenness and plant carbon fluxes from observation-constrained and modelling approaches for warm- and cold-spring years. a–f, Monthly anomalies in regionally averaged maximum composited climate (**a, d**), NDVI (**b, e**) and GPP (**c, f**) for warm- and cold-spring years, for the focus regions (**a–c**, western USA; **d–f**, Siberia). The anomalies are relative to the mean of the study period (1982–2011) and are based on maximum composites of monthly means of the seven warmest- and coldest-spring years within the study period. The observation-constrained GPP anomalies (**c, f**) stem from FluxNetG, which combined GPP estimates from flux towers with climate and satellite greenness in a machine-learning framework (see Methods). The boundaries between the climatological seasons are indicated by

vertical grey dashed lines. Uncertainty bounds (shaded areas) reflect the spread in the respective monthly anomalies within the compositing period (± 1 s.d., $n = 7$). On the basis of these anomalies, we estimate, for a warm-spring year (relative to mean conditions) in Siberia (area, $2.5 \times 10^6 \text{ km}^2$), annual GPP increases of 0.4 Pg C and 1.7 Pg C for FluxNetG and the TRENDYv6 ensemble, respectively, which corresponds to higher plant carbon uptake in the TRENDYv6 ensemble by a factor of roughly four (**f**). This is, to a large extent (about 64%), because of the overestimation of positive lagged effects in the TRENDYv6 models, but another important factor (36%) is the higher sensitivity of concurrent carbon uptake to spring warming in the TRENDYv6 models (compared to FluxNetG).

Extended Data Table 1 | Comparison of how specific processes relevant to this study are represented in the TRENDYv6 carbon-cycle models

	CLM4.5	LPX-Bern	LPI-GUESS	ISAM	CABLE	VISIT	DLEM	JSBACH	ORCHIDEE-MICT	JULES
Phenology scheme (incl. fixed leaf life span?)	Prognostic, based on carbon availability and allocation. Relatively fixed leaf lifespan (some variation in stress-deciduous PFTs)	Temperature based phenology with drought response, no fixed leaf life span for deciduous trees	Daily leaf phenology based on temperature and water availability	Carbon-gain-based dynamic phenology. The leaf onset starts when the environmental conditions are advantageous for the plant to produce leaves and to start carbon assimilation; and the leaf offset starts when the plant experiences unfavourable environmental conditions resulting in C loss, such as cold temperature, photoperiod (shorter day length), and soil moisture stress	Daily leaf phenology based on temperature and water availability for grasses and cold-deciduous woody PFTs (deciduous needle-leaf and deciduous broad-leaf) Evergreen needle-leaf trees have fixed leaf turnover rate.	Monthly leaf phenology based on temperature and water availability ¹	Satellite-constrained daily leaf phenology, no fixed leaf life span	Phenology scheme based on logistic growth with two different forest phenology types in the considered study area: evergreen and summergreen. Both types have temperature-based phenologies with different leaf shedding rates.	Leaf onset for temperate and boreal deciduous trees is driven by the accumulation of warm temperatures (growing degree day) and the number of chilling days. Leaf senescence is based on air temperature for deciduous trees and on both temperature and moisture for grasses. For all PFTs, another leaf turnover is triggered if leaf age exceeds a PFT-specific parameter of critical leaf age, and thus leaf biomass gradually decrease with time.	Daily, based on thermal time above 5°C. No drought phenology or fixed leaf age
Permafrost (yes/no?)	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No
Dynamic vegetation or static/prescribed land cover?	Prescribed land cover	Dynamic vegetation	Dynamic vegetation	Prescribed land cover	Prescribed land cover	Prescribed land cover	Prescribed land cover	Prescribed land cover	Prescribed land cover	Dynamic vegetation
Explicit representation of decid. needleleaf forests?	Yes	Yes	Yes	Yes, model explicitly account for primary and secondary decid. Needleleaf forests.	Yes	Yes (as boreal deciduous needle-leaved forest)	Yes	no (in this simulation only one PFT was used to represent temperate deciduous forests)	yes	No
Carbon allocation scheme (how and how often updated?)	Allocation to individual living tissue components updated at every model timestep (30 minutes)	Annual allocation of carbon to leaf, sapwood and root compartments, according to PFT dependent allometric rules.	annual carbon allocation from individual NPP to living tissue compartments ¹	The allocation of carbon to leaf, stem, root, and grain pools is a dynamic process based on temperature, water availability, light, and nutrients to alter the carbon allocation fractions dynamically at each model time step (1h)	Daily carbon allocation. Static fraction to roots, except during green-up of deciduous Pfts, when carbon is preferentially allocated above-ground. Relative allocation to leaves and wood is constrained by leaf-area to sapwood-area ratio.	Allocation for leaf first, then residual carbon is allocated to stem and root with a constant ratio, in order to realize the biome-specific growth form at each time step (monthly).	Daily carbon allocation from NPP to individual tissue	NPP enters the vegetation C pools daily according to fixed PFT specific proportions, adapted to adhere to structural limits.	Daily allocation from NPP to individual compartments (leaf, root, sapwood, fruit, carbohydrate reserve), according to PFT-specific fractions and regulated by light limitation, soil moisture and soil temperature.	10 daily allocation between growth and spreading. Growth allocation is split between leaves, roots and stem. Spreading allocation to increasing fractional coverage of pools.

PFT, plant functional type; NPP, net primary productivity.

An abstract drawing from the 73,000-year-old levels at Blombos Cave, South Africa

Christopher S. Henshilwood^{1,2*}, Francesco d'Errico^{1,3}, Karen L. van Niekerk¹, Laure Dayet^{3,4}, Alain Queffelec³ & Luca Pollarolo^{5,6}

Abstract and depictive representations produced by drawing—known from Europe, Africa and Southeast Asia after 40,000 years ago—are a prime indicator of modern cognition and behaviour¹. Here we report a cross-hatched pattern drawn with an ochre crayon on a ground silcrete flake recovered from approximately 73,000-year-old Middle Stone Age levels at Blombos Cave, South Africa. Our microscopic and chemical analyses of the pattern confirm that red ochre pigment was intentionally applied to the flake with an ochre crayon. The object comes from a level associated with stone tools of the Still Bay techno-complex that has previously yielded shell beads, cross-hatched engravings on ochre pieces and a variety of innovative technologies^{2–5}. This notable discovery predates the earliest previously known abstract and figurative drawings by at least 30,000 years. This drawing demonstrates the ability of

early *Homo sapiens* in southern Africa to produce graphic designs on various media using different techniques.

Blombos Cave (BBC) is situated on the southern Cape coast, about 300 km east of Cape Town (34° 24' 51" S, 21° 13' 19" E). Excavations commenced in 1991 and are on-going. The site contains well-stratified Middle Stone Age (MSA) deposits dating to between 100 and 72 thousand years ago (ka)⁶, topped by a layer of sterile aeolian dune sand (dated to 70 ka) and Later Stone Age layers dated to 2 ka (Fig. 1). The MSA sequence consists of four phases, of which the upper two—'BBC M1' and 'BBC M2 upper'—are associated with the Still Bay techno-complex, dated to about 77–73 ka⁷. These phases contain bifacial foliate points, of which 12% were heated before final shaping using pressure flaking^{5,8}. Other cultural markers of the Still Bay from these layers include bone awls and spear points⁴, possible engravings of parallel and joining lines

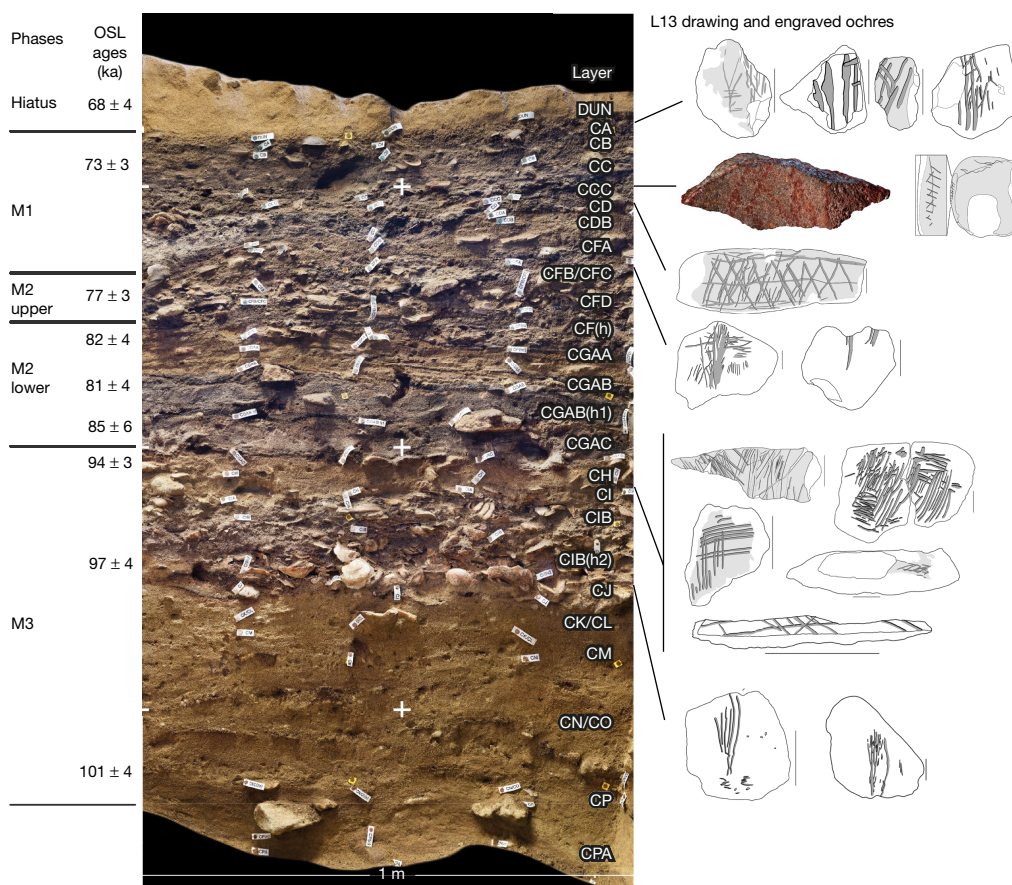


Fig. 1 | Stratigraphy of the south section of Blombos Cave. Left, phases and optically stimulated luminescence dates for the Middle Stone Age levels at Blombos Cave. Centre, labels for individual layers superimposed

on section. Right, layers from which the L13 silcrete flake with ochre drawings and previously described engraved ochre pieces were recovered. Scale bars, 1 cm.

¹SFF Centre for Early Sapiens Behaviour (SapienCE), University of Bergen, Bergen, Norway. ²Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa. ³CNRS UMR 5199, University of Bordeaux, Bordeaux, France. ⁴Laboratoire TRACES UMR 5608, Université Toulouse Jean Jaures, Toulouse, France. ⁵Unité d'Anthropologie/Laboratoire Archéologie et Peuplement de l'Afrique, Geneva, Switzerland. ⁶School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa. *e-mail: christopher.henshilwood@uib.no



Fig. 2 | Image of the Blombos Cave silcrete flake L13 displaying the drawn lines that form a cross-hatched pattern. Image credit, C. Foster.

on bone^{4,9}, beads made from *Nassarius kraussianus* shells (67 recovered thus far, some of which are stained with ochre^{3,10}) and pieces of ochre engraved with geometric patterns, eight of which have been recovered thus far—of which two display distinct cross-hatched designs^{2,11}.

The 'M2 lower' phase (dated to about 85–82 ka) is a low-intensity occupational horizon with no Still Bay type artefacts or engraved ochre. The M3 phase (dated to about 101–94 ka) contains abundant natural and used ochre, as well as ten pieces of ochre engraved with geometric designs that include three crosshatched motifs¹¹. In addition, an in situ toolkit consisting of ochre, heated seal bone, charcoal and associated processing materials—used to create a liquid pigmented compound stored in abalone shells—was recovered from a layer of this phase that dates to about 100 ka⁶.

The silcrete flake with a cross-hatched pattern described here was excavated in 2011 and comes from the M1 phase (layer CCC, square G7b). The objects recovered from this layer were gently washed with running tap water in the field and the laboratory, and dried at air temperature. The pattern on the piece was noticed during analysis of the lithic debitage. The piece was numbered L13 in the laboratory (Fig. 2, Supplementary Data and Supplementary Video). The flake is coarse-grained silcrete (length 38.6 mm, width 12.8 mm and height 15.4 mm). The pattern is on a slightly concave smooth face—the flake's striking platform—and extends to a flake scar on the same surface. It consists of a set of six straight sub-parallel lines (Fig. 3, lines 1–6) crossed obliquely by three slightly curved lines (lines 7–9). Line 6 partially overlaps the edge of the flake scar, which suggests it was made after the flake became detached. All the lines are discontinuous and one (line 5) is wider and better defined than the others. Microscopic and chemical analyses confirm that the lines on L13 are composed of haematite-rich powder, commonly called ochre (Extended Data Fig. 1, Supplementary Information and Supplementary Table 2). They differ markedly in colour and composition from the silcrete and the orange calcite patches present on the same surface. The abrupt termination of all lines on the fragment edges indicates that the pattern originally extended over a larger surface (Fig. 3b)—the pattern was probably more complex and structured in its entirety than in this truncated form.

Experimentally marking silcrete flakes with ochre crayons or paint indicates that the lines on L13 were produced with a crayon and thus constitute a drawing (Fig. 4, Extended Data Figs. 3, 4, Supplementary Information and Supplementary Table 1). Discontinuous lines consisting of loose powder in recesses and firmly adhering ochreous patches on elevated areas are produced when experimentally marking a silcrete flake with an ochre crayon. Parallel striations are visible on the patches. These same diagnostic features are seen on the lines composing the pattern on L13. By contrast, our experimental painted lines have clear edges, are solidly filled and show no striations (Extended Data Fig. 7a). The deposits produced by the two techniques are still clearly visible on the experimental material after rinsing under running water (Extended Data Fig. 7c).

The width of the lines on the archaeological material, compared to their experimental counterpart, indicates that they were probably drawn with a pointed ochre crayon (Extended Data Fig. 5). Lines 2, 3, 4, 7, 8 and 9 were made by a single stroke with a 1.3–3.3-mm-wide ochre tip. Line 5 is wider and the ochre is more continuous, probably because it was created with multiple strokes. On experimental drawn lines, loose powder followed by a compacted streak is an indicator of

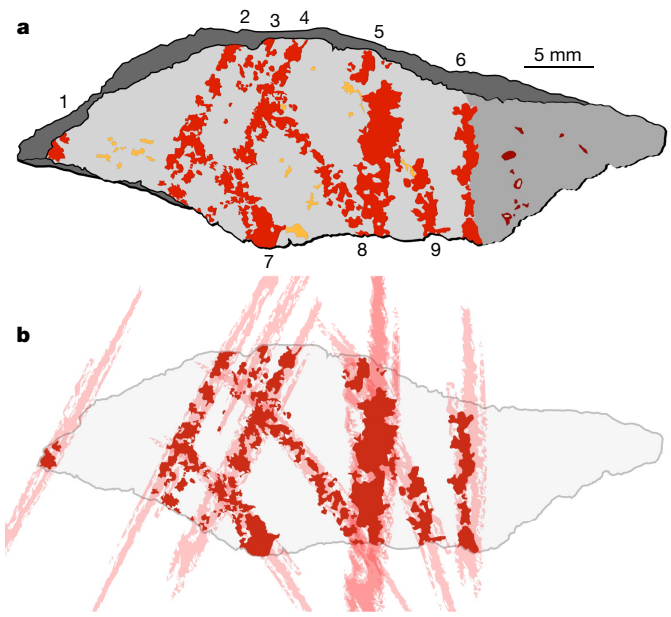


Fig. 3 | Renditions of L13. a, Tracing with the drawn red lines numbered, the calcite patches shown in orange and the ochre residues in dark red. The ground surface is coloured light grey, the darker grey area indicates a flake scar and the darkest grey indicates the breakage fractures after L13 became detached from the originally larger grindstone. **b,** Schematic of lines extending beyond the outline of the present flake.

stroke direction. When applied to the L13 drawing, this observation shows that the single-stroke lines 2, 3 and 4 were drawn in the same direction (Extended Data Figs. 3, 4). Line 5 is a multiple-stroke line, perhaps produced by a to-and-fro motion (Fig. 3). Line 6 does not show clear indication of directionality. Line 8—and probably lines 7 and 9—was drawn in a direction opposite to that of lines 2, 3 and 4, which could indicate that the object was turned during drawing. The order in which the lines were drawn could not be established.

The surface on L13 with the drawing is smoothed by grinding and shows microscopic haematite-rich residues, which indicates that the object is a flake from a grindstone that was used to process ochre before the drawing was made (Extended Data Fig. 2). Subsequent cleaning of the grindstone by the inhabitants of BBC removed most traces of loose ochre powder (produced during grinding), and left a surface that was almost clean but which retained minute traces of ochre. Tribological analysis conducted with two different methods supports the grindstone hypothesis; it shows that the surface with the drawing is significantly smoother than the other faces of L13 and typical cortical and knapped surfaces of other pieces of silcrete recovered from MSA levels at BBC, which confirms that the smoothing recorded on L13 cannot be due to natural processes (Extended Data Fig. 6a, b, Supplementary Information and Supplementary Table 3).

Experimental reproduction of the lines found on L13 using the same technique (that is, using a pointed ochre crayon) indicates that the drawing was more visible when it was produced and that the loose powder that originally composed the lines was subsequently lost through taphonomic processes and rinsing. The lines produced on smooth silcrete—such as on L13—are better defined than on rough surfaces; and the ochre crayon used for the design was probably soft and produced lines that adhered well to the silcrete. We conclude that the ochre crayon was intentionally used to produce a cross-hatched design.

Ochre powder was used for practical purposes by MSA people; for example, as a glue additive and perhaps as a sun screen^{12,13}. Experimental reproduction of numerous lines longer than all those on L13 (Extended Data Fig. 3b, Supplementary Information) produced less than 1 mg of ochre powder, thus discounting a utilitarian objective for the production of the L13 lines.

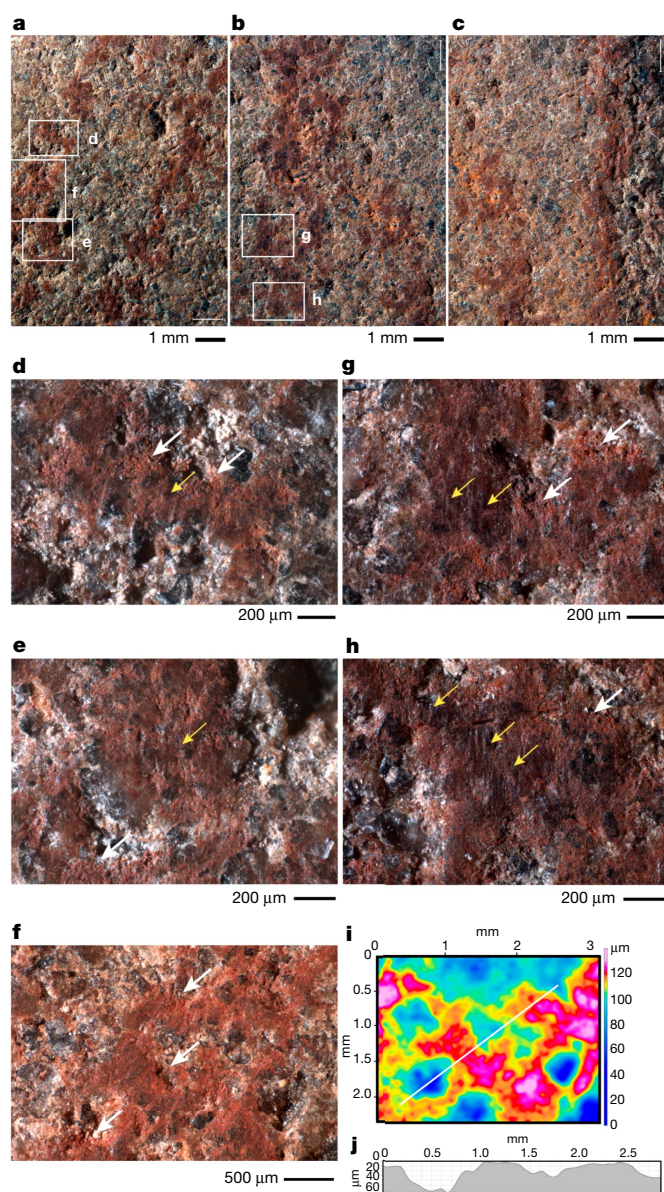


Fig. 4 | Close-up views of the drawn lines on the L13 surface. a, Middle portions of lines 3 and 4 with location of the micrographs shown in d–f. b, Lines 8 and 9 on the proximal portion of the flake, with location of the micrographs g, h. c, Lines 9 and 6 on the proximal portion of the flake. i, Depth map of micrograph shown in f, with the location of the section shown in j indicated by a white line. Large white large arrows (d–h) point to deposits of powdery ochre preserved in recesses, small yellow arrows (d, e, g, h) point to prominent areas with compacted ochre deposits on which remnants of striations are still visible. These features are similar to those produced when a silcrete flake is experimentally marked with an ochre crayon (Extended Data Fig. 3b).

Abstract engravings are known from several archaeological sites that pre-date the Later Stone Age (which began about 42–44 ka) in Africa, and from the Upper Palaeolithic (dated to about 44 ka) in Europe; these include an engraved shell from Trinil dated to 540 ka¹⁴, an engraved bone from Bilzingsleben at 370 ka¹⁵, engraved ochre from BBC (100–73 ka)¹¹, engraved cortices from Qafzeh (90 ka)¹⁶ and Quneitra (60 ka)¹⁷, engraved ostrich egg shells from Diepkloof¹⁸ and Klipdrift Shelter (65–59 ka)¹⁹ and engraved bedrock from Gorham's Cave (over 40 ka)²⁰. A date of 66–64 ka—which would indicate Neanderthal authorship—has recently been proposed for ochre markings from three caves in Spain^{21,22}. However, the earliest previously known evidence for drawing techniques comes from much younger

sites that post-date 42 ka, such as Chauvet^{23,24}, El Castillo²⁵, Apollo 11²⁶ and Maros caves²⁷.

The discovery of L13 demonstrates that drawing was part of the behavioural repertoire of populations of early *Homo sapiens* in southern Africa at about 73 ka. It demonstrates their ability to apply similar graphic designs on various media using different techniques. The discovery of abstract engravings on ochre, with patterns comparable to L13, from levels at BBC dated to 100–73 ka (Fig. 1) and the production of an ochre-rich paint stored in abalone shells⁶ suggest that drawings and possibly paintings may have been produced in older MSA levels, perhaps since 100 ka. The cross-hatched pattern of L13 pre-dates by at least 30,000 years the earliest previously known abstract and figurative drawings. This finding supplements previous evidence reflecting cultural modernity and symbol use that has already been identified in the MSA levels at BBC through the discovery of personal ornaments, elaborate bone tools, engravings, and the production and storage of pigmented compounds. The L13 drawing adds a further dimension to our understanding of the processes that shaped the behaviour and cognition of early *H. sapiens*.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0514-3>.

Received: 16 February 2018; Accepted: 7 August 2018;

Published online 12 September 2018.

- d'Errico, F. & Stringer, C. B. Evolution, revolution or saltation scenario for the emergence of modern cultures? *Phil. Trans. R. Soc. Lond. B* **366**, 1060–1069 (2011).
- Henshilwood, C. S. et al. Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science* **295**, 1278–1280 (2002).
- Henshilwood, C., d'Errico, F., Vanhaeren, M., van Niekerk, K. & Jacobs, Z. Middle Stone Age shell beads from South Africa. *Science* **304**, 404 (2004).
- d'Errico, F. & Henshilwood, C. S. Additional evidence for bone technology in the southern African Middle Stone Age. *J. Hum. Evol.* **52**, 142–163 (2007).
- Mourre, V., Villa, P. & Henshilwood, C. S. Early use of pressure flaking on lithic artifacts at Blombos Cave, South Africa. *Science* **330**, 659–662 (2010).
- Henshilwood, C. S. et al. A 100,000-year-old ochre-processing workshop at Blombos Cave, South Africa. *Science* **334**, 219–222 (2011).
- Jacobs, Z., Hayes, E. H., Roberts, R. G., Galbraith, R. F. & Henshilwood, C. S. An improved OSL chronology for the Still Bay layers at Blombos Cave, South Africa: further tests of single-grain dating procedures and a re-evaluation of the timing of the Still Bay industry across southern Africa. *J. Archaeol. Sci.* **40**, 579–594 (2013).
- Villa, P., Soressi, M., Henshilwood, C. S. & Mourre, V. The Still Bay points of Blombos Cave (South Africa). *J. Archaeol. Sci.* **36**, 441–460 (2009).
- d'Errico, F., Henshilwood, C. S. & Nilssen, P. An engraved bone fragment from c. 70,000-year-old Middle Stone Age levels at Blombos Cave, South Africa: implications for the origin of symbolism and language. *Antiquity* **75**, 309–318 (2001).
- Vanhaeren, M., d'Errico, F., van Niekerk, K. L., Henshilwood, C. S. & Erasmus, R. M. Thinking strings: additional evidence for personal ornament use in the Middle Stone Age at Blombos Cave, South Africa. *J. Hum. Evol.* **64**, 500–517 (2013).
- Henshilwood, C. S., d'Errico, F. & Watts, I. Engraved ochres from the Middle Stone Age levels at Blombos Cave, South Africa. *J. Hum. Evol.* **57**, 27–47 (2009).
- Wadley, L. Ochre crayons or waste products? Replications compared with MSA 'crayons' from Sibudu cave, South Africa. *Before Farming* **2005**, 1–12 (2005).
- Rifkin, R. F. et al. Evaluating the photoprotective effects of ochre on human skin by in vivo SPF assessment: implications for human evolution, adaptation and dispersal. *PLoS ONE* **10**, e0136090 (2015).
- Joordens, J. C. et al. *Homo erectus* at Trinil on Java used shells for tool production and engraving. *Nature* **518**, 228–231 (2015).
- Mania, D. & Mania, U. Deliberate engravings on bone artefacts of *Homo erectus*. *Rock Art Res.* **5**, 91–107 (1988).
- Hovers, E., Vandermeersch, B. & Bar-Yosef, O. A Middle Palaeolithic engraved artefact from Qafzeh Cave, Israel. *Rock Art Res.* **14**, 79–87 (1997).
- Marshack, A. A Middle Paleolithic symbolic composition from the Golan heights: the earliest known depictive image. *Curr. Anthropol.* **37**, 357–365 (1996).
- Texier, P.-J. et al. The context, form and significance of the MSA engraved ostrich eggshell collection from Diepkloof Rock Shelter, Western Cape, South Africa. *J. Archaeol. Sci.* **40**, 3412–3431 (2013).
- Henshilwood, C. S. et al. Klipdrift Shelter, southern Cape, South Africa: preliminary report on the Howiesons Poort layers. *J. Archaeol. Sci.* **45**, 284–303 (2014).
- Rodríguez-Vidal, J. et al. A rock engraving made by Neanderthals in Gibraltar. *Proc. Natl Acad. Sci. USA* **111**, 13301–13306 (2014).
- Hoffmann, D. L. et al. U-Th dating of carbonate crusts reveals Neanderthal origin of Iberian cave art. *Science* **359**, 912–915 (2018).

22. Pearce, D. G. & Bonneau, A. Trouble on the dating scene. *Nat. Ecol. Evol.* **2**, 925–926 (2018).
23. Quiles, A. et al. A high-precision chronological model for the decorated Upper Paleolithic cave of Chauvet-Pont d'Arc, Ardèche, France. *Proc. Natl Acad. Sci. USA* **113**, 4670–4675 (2016).
24. Sadier, B. et al. Further constraints on the Chauvet cave artwork elaboration. *Proc. Natl Acad. Sci. USA* **109**, 8002–8006 (2012).
25. Pike, A. W. et al. U-series dating of Paleolithic art in 11 caves in Spain. *Science* **336**, 1409–1413 (2012).
26. Wendt, W. E. 'Art mobilier' from the Apollo 11 Cave, South West Africa: Africa's oldest dated works of art. *S. Afr. Archaeol. Bull.* **31**, 5–11 (1976).
27. Aubert, M. et al. Pleistocene cave art from Sulawesi, Indonesia. *Nature* **514**, 223–227 (2014).

Acknowledgements Partial funding for this research was provided to C.S.H., K.L.v.N. and F.d'E. by the Research Council of Norway through its Centres of Excellence funding scheme, Centre for Early Sapiens Behaviour (SapienCE), project number 262618; to C.S.H. by a South African National Research Foundation Research Chair (SARCH) at the University of the Witwatersrand and the Evolutionary Studies Institute at the University of the Witwatersrand, and the University of Bergen, Norway; F.d'E., L.D. and A.Q. by the LaScArBx, a research programme supported by the ANR (ANR-10-LABX-52). We thank C. Foster for the image in Fig. 2; P. Keene for assistance in the Cape Town laboratory, I. Svahn for assistance with electron microscopy in Bordeaux, G. Devilder for his input on Fig. 3 and M. Haaland for his stratigraphy image on Fig. 1.

Reviewer information Nature thanks J. C. A. Joordens, G. van den Bergh and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.S.H. and K.L.v.N. directed the excavations at Blombos Cave. C.S.H., F.d'E. and K.L.v.N. planned the methodology for examination of L13, and conceived and carried out the experimental replication tests. F.d'E. and L.D. carried out the microscopic analysis of L13 and experimental lines. L.D. carried out the chemical analyses of L13. A.Q. carried out the tribological analysis of the surfaces of L13 and produced the MP4 video (Supplementary Video) and three-dimensional PDF (Supplementary Data) of L13. L.P. recovered L13 during lithic analysis and recognized its importance. C.S.H., F.d'E., K.L.v.N., L.D. and A.Q. co-wrote the paper. L.P. contributed to editing the final paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0514-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0514-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.S.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

L13 was photographed using various macro-lenses, and examined and photographed with a motorized Leica Z6 APOA equipped with a DFC420 digital camera linked to LAS Montage and Leica Map DCM 3D computer software. Sections and three-dimensional models of selected portions of the lines on L13 were obtained with the LAS Montage, or by exporting depth maps obtained with the LAS Montage into the Leica Map DCM 3D. An image of the surface with the cross-hatching was obtained by importing overlapping micrographs from the LAS Montage into Adobe Illustrator and then producing a tracing of the red patches and other features identified on L13. This tracing was compared to the original under a microscope and corrected as required. Data on the morphology, size, number of patches composing the lines on L13 and presence of striations on these lines were recorded.

Unmodified pieces of ochre (Extended Data Fig. 7a) with narrow pointed or linear edges and variable texture and hardness were used to experimentally mark silcrete flakes (Extended Data Fig. 7b). These flakes are archaeological objects that derive from weathered and eroding MSA levels on the BBC talus.

The silcrete flakes were carefully cleaned with a brush under running water between marking experiments to remove all traces of ochre from their surfaces. In the first experiment, both pointed and linear ochre pieces were used to produce single- and multiple-stroke straight and curved lines. Multiple strokes were produced by repeatedly passing the ochre edge over the silcrete surface in the same direction or with a to-and-fro movement in an effort to accurately superimpose each new line on those previously made. In a second experiment, four pointed and four linear ochre edges were used to produce a sequence of six single-stroke, parallel, straight four-centimetre-long lines per edge (Supplementary Table 1).

The width and length of the microfacets created on the ochre pieces by the marking process and the maximum and minimum width of the lines produced on the silcrete flakes were measured with a digital calliper after each line was produced. In a third experiment, fine-grained ochre powder, produced by rubbing ochre pieces on a grindstone, was mixed with water in three different concentrations to produce three gradations of viscosity (thin, medium and thick).

Wooden sticks of 1 mm and 2 mm in diameter were gently crushed at one end with a hammer over a length of 1 cm to create a brush, and this was used to apply the three types of paint onto flat surfaces on the silcrete flakes.

The experimental lines made with the ochre pieces and wooden applicators were examined and photographed with the same equipment and procedures used for examining L13 immediately after their production and after being gently washed for 10 s under running tap water and dried at air temperature. This is the same cleaning process used for L13 and other BBC lithics recovered during sieving. When examining the experimental drawn and painted lines under the microscope, particular attention was paid to identifying features diagnostic of the marking technique, the direction in which the lines were made, the type of ochre edge used and the way in which washing altered the lines. In a final experiment, loose ochre powder produced by drawing nine four-centimetre-long straight lines with pointed and linear pieces on a silcrete flake was recovered and weighed on a scale with an accuracy of 1 mg.

Scanning electron microscopy with energy-dispersive X-ray spectroscopy (SEM–EDS) analyses were performed using a FEI Quanta 200. Back-scattered electron images (BSE) and elemental analyses were conducted under a low

vacuum mode with an accelerating voltage of 15 kV. BSE images were produced with a SiLi detector and EDS analyses with a SDD-EDAX detector. The EDS analyses were conducted under similar magnifications ($\times 50$, $\times 100$ or $\times 200$, and $\times 500$ or $\times 1,000$), at the same working distance (10 mm) and with the same acquisition time (100 s) for each EDS spectrum. Semi-quantitative data were calculated in weight percentages and normalized to 100%, C and O being included. Major ($>10\%$), minor ($>3\%$) and minor-to-trace elements ($<3\%$) were distinguished²⁸. For Raman analysis we used a SENTERRA dispersive Raman microscope (Bruker), equipped with an internal calibration system. The analyses were done with a 785-nm laser and a power of 1 mW to avoid transformation of mineral phases. Acquisition time was set to between 30 s and 70 s, and several co-additions of the signal if necessary. The working area was observed through the integrated colour camera, and data were collected with the software package OPUS 7.2.

High-resolution surface topography was acquired with a Sensofar S neox confocal microscope driven by SensoScan 6 software (Sensofar). For this study, we selected thirty 1.67×1.25 -mm areas that were measured with a $20\times$ objective (NA 0.45). This allowed for a spatial resolution of $0.65 \mu\text{m}$ and a vertical resolution of $0.31 \mu\text{m}$. Seven areas were measured outside of the red lines on the surface with the drawing, seven on the other surfaces of L13, ten on the knapped surfaces and ten on the natural cortical surfaces of two fine-grained silcrete flakes from the MSA levels at BBC (Extended Data Fig. 6a). Three-dimensional reconstructions of the measured areas were visually compared. The location of the measured areas on L13 was randomly selected (Extended Data Fig. 6c). Tribological analysis of these surfaces entailed pre-treatments and the calculation of surface texture parameters according to the ISO standard 25178, using the SensoMap software. Pre-treatments conducted on the raw acquisitions consisted of (1) filling non-measured points using the neighbourhood valid points algorithm, (2) removing the general shape of the surface by subtracting a second-degree polynomial, (3) separating waviness from roughness by the application of a Gaussian filter with a 0.25 -mm cut-off value, so that only roughness is kept for the calculation of surface texture parameters. Height, functional, spatial, hybrid and functional volume parameters were calculated for each quarter of each pre-treated area. Kruskal–Wallis multiple comparison tests²⁹ were applied to the dataset.

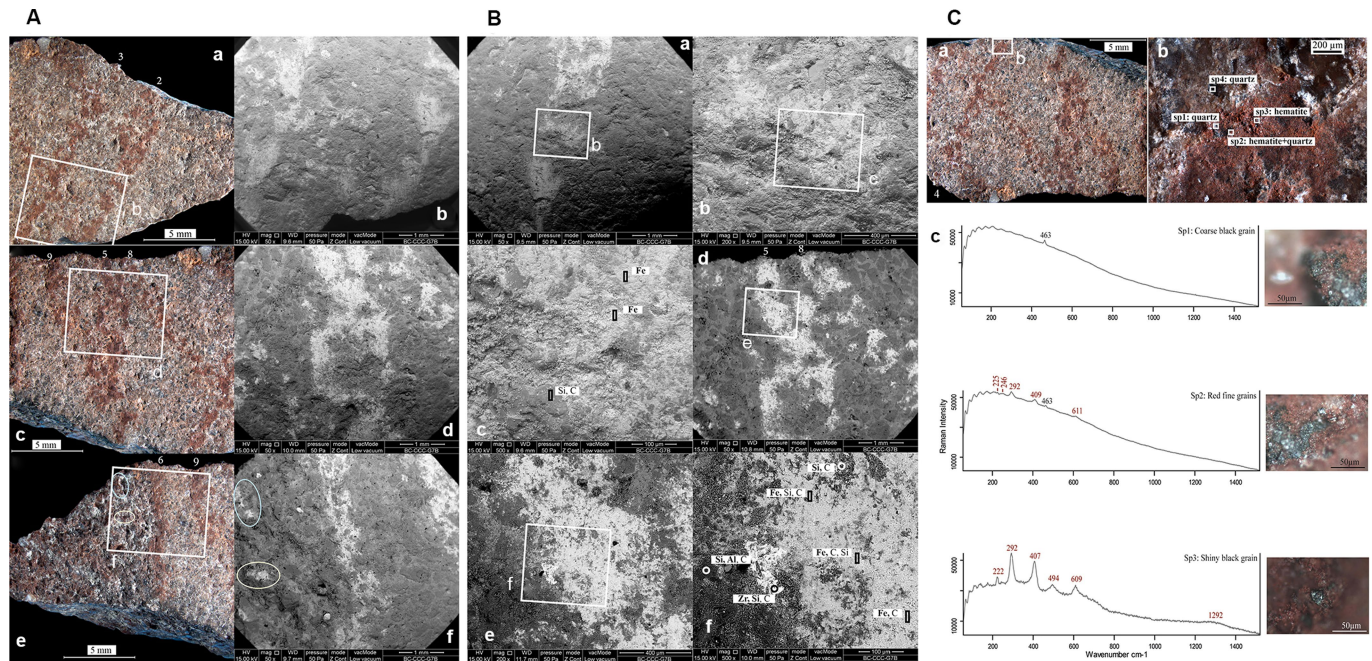
Scale sensitive fractal analysis using SensoMap 7.4.8443 software was also applied to L13³⁰. This analysis calculates the difference between the topographic surface and its planimetric area. The calculation is made at many consecutive scales by reducing the area of the triangles tiling the surface. Results are represented by an s-curve in which the rising indicates the scale at which the surface becomes rough.

A high-resolution three-dimensional model was created by photogrammetry using the Photoscan Standard software (Agisoft, version 1.4.0). Fifty-six photos taken with a macro lens from different points of view were processed to create the model and apply the texture (Supplementary Data and Supplementary Video).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

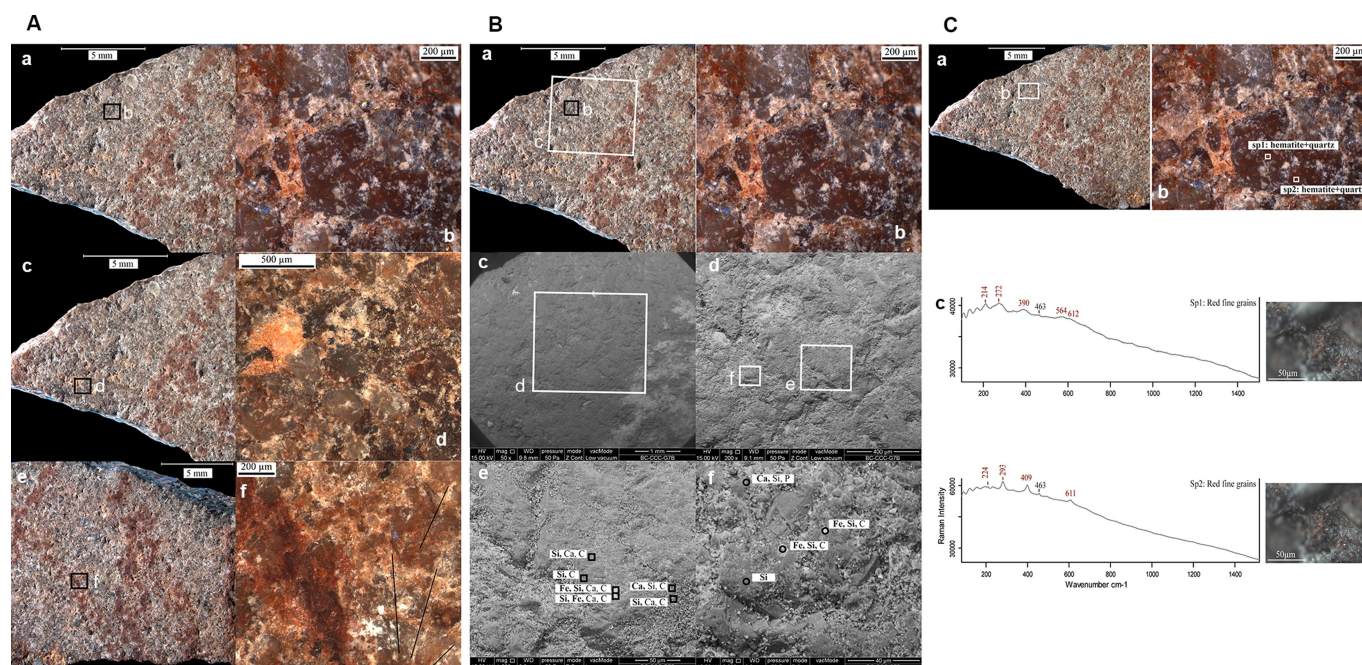
Data availability. All data generated or analysed during this study are included in the published article and its Supplementary Information.

28. d'Errico, F. et al. The technology of the earliest European cave paintings: El Castillo Cave, Spain. *J. Archaeol. Sci.* **70**, 48–65 (2016).
29. Siegel, S. & Castellan, N. J. Jr. *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1988).
30. Brown, C. A. in *Characterisation of Areal Surface Texture* (ed. Leach, R.) 129–153 (Springer, Berlin, 2013).



Extended Data Fig. 1 | Microscopic examination and chemical analyses of the juxtaposed patches of red deposit that form the drawn lines on L13 and the smoothed surface of the silcrete flake. The lines consist mainly of fine-grained iron oxide (Fe), that were applied to the surface, as no haematite occur naturally in the silcrete raw material of L13. **A**, Photographs (left) and SEM-EDS images (right) of the red lines of the surface of L13. In the subpanels of **A**, images in **a**, **b** show lines 2 and 3; **c**, **d** show lines 5, 8 and 9; and **e**, **f** show lines 6 and 9 and red spots on a flake scar. The white rectangles in **a**, **c** and **e** indicate the areas that are enlarged in subpanels **b**, **d** and **f**, respectively. Notice the white appearance of the lines in the back-scattered electrons SEM-EDS images due to

the presence of iron rich deposits. **B**, SEM-EDS images (back-scattered electrons) of line 2 and 5. Subpanels **a**–**c** show line 2; **d**–**f** show line 5. White squares indicate areas that are enlarged in the image with the corresponding letter. The rectangles and black/white circles in subpanels **c**, **f** show differences in elemental composition between the drawn lines (light areas) and the silcrete surface (dark areas). **C**, Raman analysis of line 4. Subpanel **a** shows a photograph with the location of the analysed area (white rectangle). Subpanel **b** shows the analysed spots and identified minerals. Subpanel **c** shows Raman spectra and micrographs of the analysed areas with peaks identifying haematite (red numbers) and quartz (black numbers).

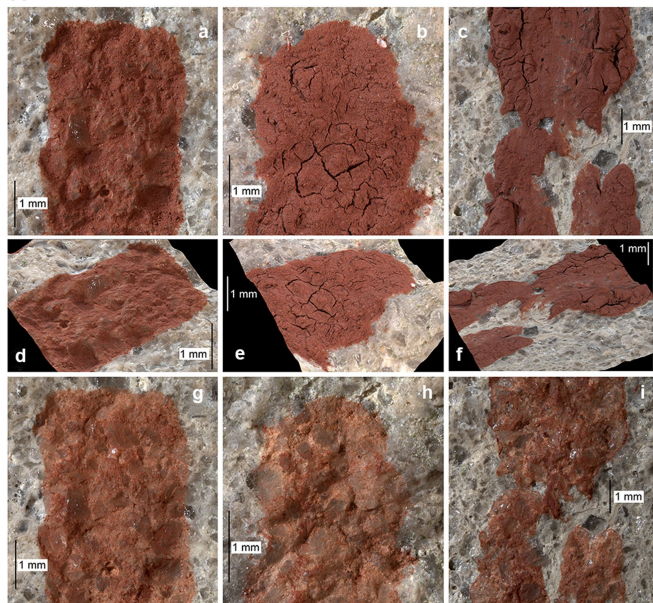


Extended Data Fig. 2 | Microscopic examination and chemical analyses of microresidues. The microresidues on the smoothed silcrete surface outside of the lines differ in Fe content from the red lines, which—along with the presence of microstriations—supports the theory that the silcrete flake was part of an ochre grindstone before the drawing was made.

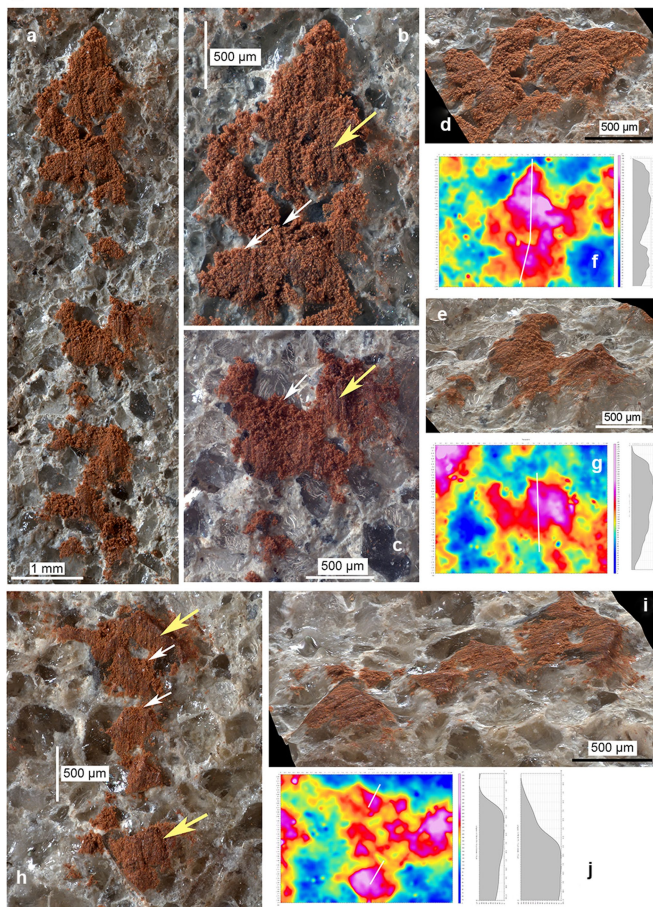
A, Photographs and micrographs of the lines drawn on L13. Black squares in subpanels a, c, e indicate the areas enlarged in the adjacent subpanels b, d, f. Red residues are clearly visible on the matrix and on quartz grains. f, Black lines highlight superficial randomly oriented striations. **B,** SEM-EDS analysis of the silcrete outside the drawn lines. In subpanels a, c, d, black and white squares indicate the areas enlarged in the adjacent

photograph. In subpanels e, f, the analysed spots (black squares and circles) identify the presence of isolated iron-rich particles on the surface of the matrix and the quartz grains. **C,** Raman analysis of microresidues preserved in quartz grain pits. Subpanel a shows a photograph with the location of the analysed area (white rectangle). Subpanel b shows analysed spots (white squares) and identified minerals. Subpanel c shows Raman spectra and micrographs of the analysed areas, with peaks identifying haematite (red numbers) and quartz (black numbers). The area shown in subpanel b of panel C is the same as the area shown in subpanel b of panel B.

A



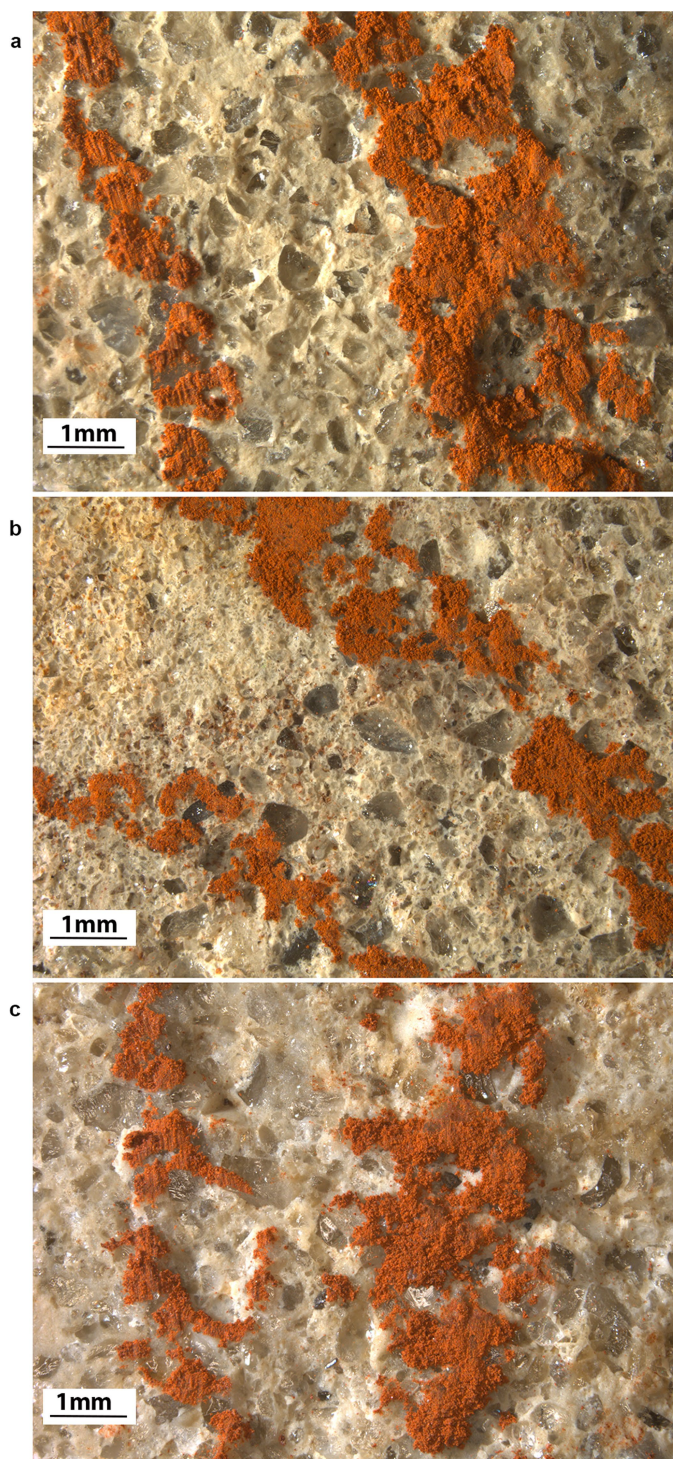
B



Extended Data Fig. 3 | See next page for caption.

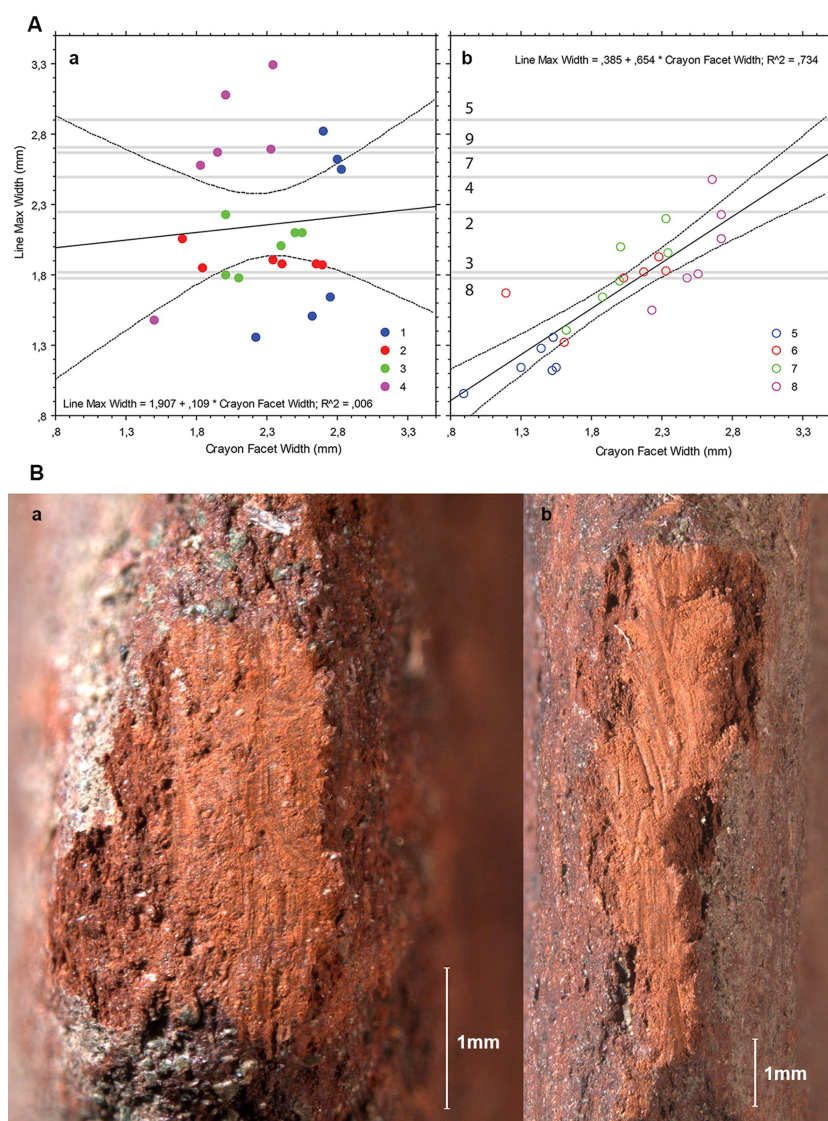
Extended Data Fig. 3 | Results from experimental marking of silcrete surfaces with ochre paint of varying viscosities and with an ochre crayon, and subsequent rinsing. **A**, Micrographs of experimentally painted lines before and after rinsing. Subpanels **a–c** show lines produced by applying a liquid (**a**), viscous (**b**) and very viscous (**c**) paint with a thin wooden brush on a silcrete surface. Subpanels **d–f** show the three-dimensional rendering of the same lines showing the surface topography. Subpanels **g–i** show the same lines after gently rinsing the surface of the silcrete under running tap water. **B**, Lines produced experimentally on a silcrete flake with an ochre crayon. Subpanel **a** shows a single-stroke line drawn from the top to the bottom. Subpanels **b, c** show close-up views and three-dimensional renderings (subpanels **d, e**) of selected areas of

subpanel **a**. Subpanels **h, i** show a photograph (**h**) and three-dimensional rendering (**i**) of a single-stroke line produced from the top to the bottom after gently rinsing the silcrete flake under running tap water. Subpanels **f, g** and **j** are depth maps and sections of **b, c** and a selected area of **h**, respectively. The locations of the sections are indicated on the depth maps by white bars. White arrows indicate deposits of powdery ochre preserved in recesses, and larger yellow arrows indicate prominent areas with compacted ochre deposits covered by striations. Compacted patches of ochre covered by striations and small deposits of ochre powder in recesses are preserved after the rinsing; these features and lines are similar to those on L13 (Fig. 4).



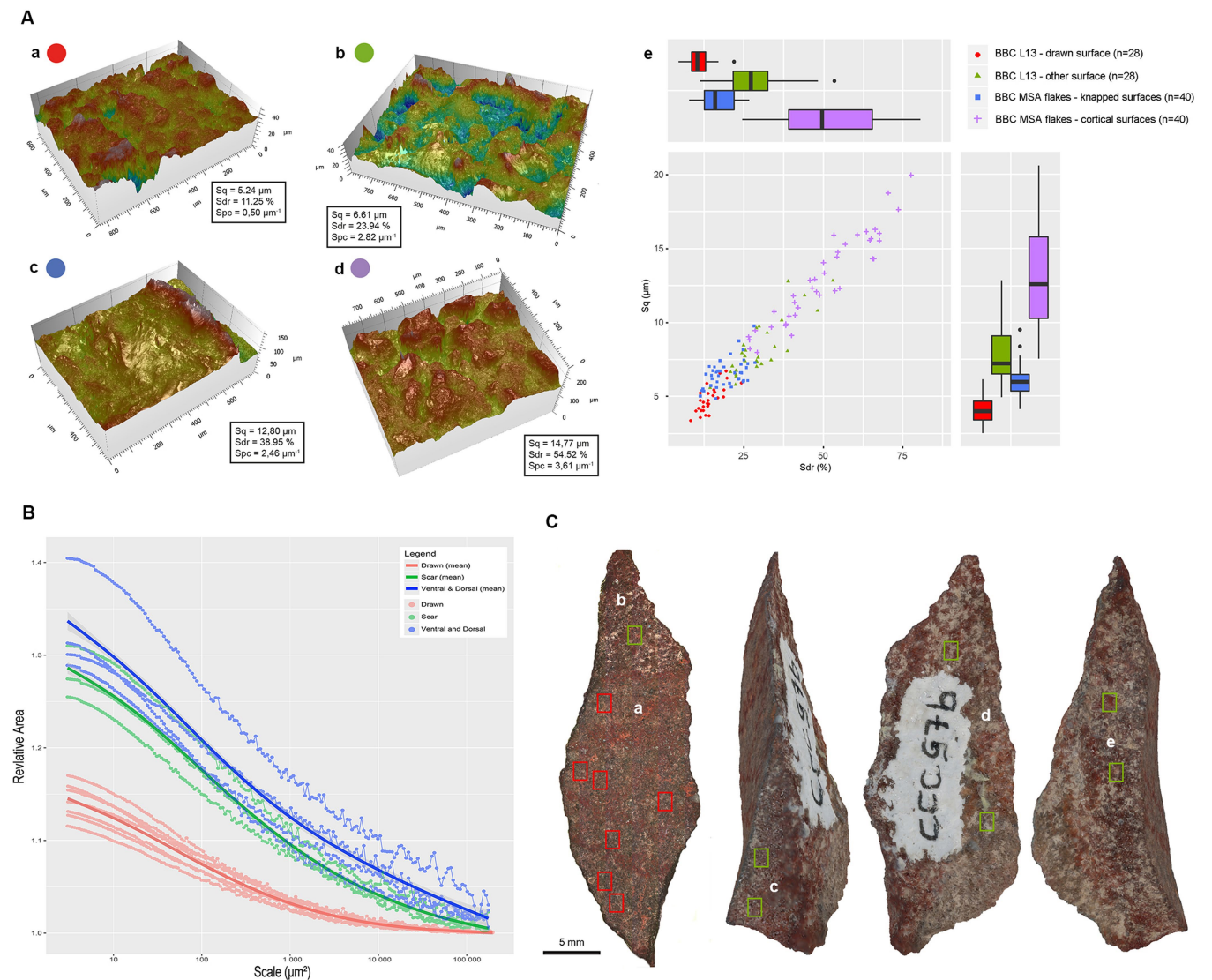
Extended Data Fig. 4 | Lines produced experimentally on silcrete flakes.

These images show that, as it is difficult to exactly superimpose a new line on a previous one, superimposing a line on a previous line generally results in a wider line. Unidirectional, superimposed lines retain the same features observed on a single-stroke line. Multiple lines produced by a to-and-fro movement of the ochre edge show microscopic evidence that the crayon was moved in both directions. **a**, Straight single- (left) and five-stroke line (right) produced from the top to the bottom. **b**, Curved single- (left) and five-stroke line (right) produced from top left to the bottom right. **c**, Straight single- (left) and five-stroke line (right) produced by a to-and-fro motion. The lines in this figure were not rinsed with water.



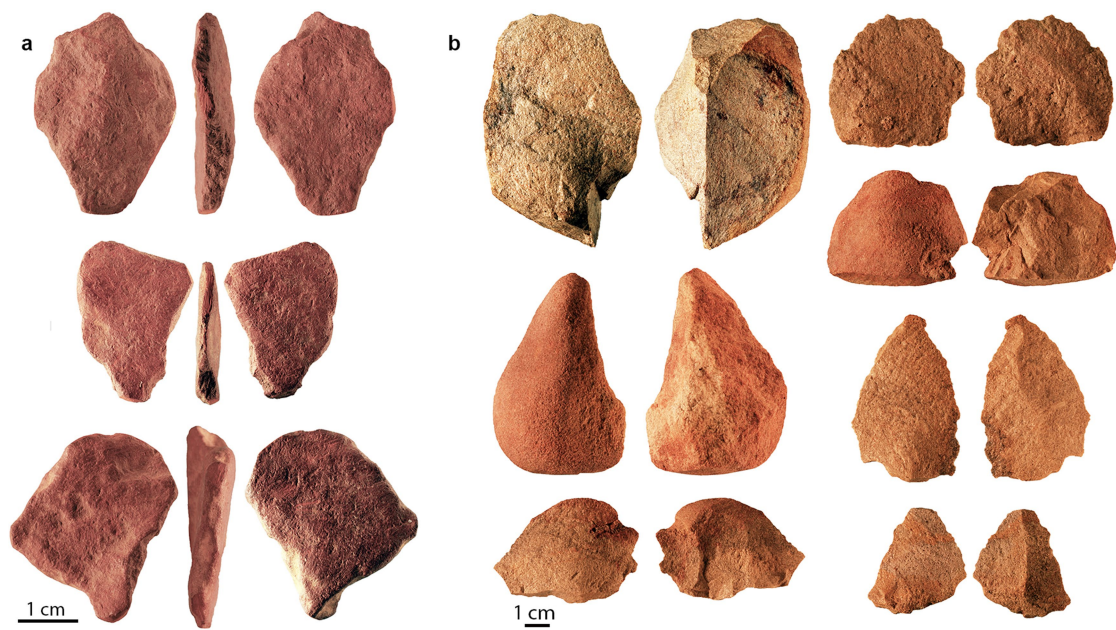
Extended Data Fig. 5 | Experimental marking of silcrete flakes with a variety of ochre crayons. The morphology of lines will depend on the properties and composition of the ochre, the roughness of the silcrete surface, the pressure exerted and the morphology of the ochre area in contact with the silcrete. In general, soft, plastic, clayish ochre will produce thicker and more continuous lines than silty or sand-rich ochre. Lines on fine-grained silcrete will be better defined than those on coarse silcrete. Stronger pressure will produce comparatively wider, thicker and better defined lines. Six lines made with each of eight unmodified ochre crayons had a maximum width ranging from about 0.9 to 3.3 mm. Lines produced with a pointed ochre crayon tend to be wider and more variable in width than those made with a linear edge. The width of lines made with a linear edge is strongly correlated with the maximum width of the facets on the ochre piece. By contrast, no correlation is observed between the lines

made with pointed crayons and the maximum width of the facets on the crayon. The width of the lines on the drawn cross-hatching present on L13 is comparable with that of the experimental lines. The range (1.8–2.9 mm) of this width best fits the width variability observed when marking the silcrete with a pointed crayon rather than an edge. This indicates that a pointed ochre crayon was used to produce the cross-hatching and that the facet of the crayon in contact with the silcrete was about 1.3–2.9 mm wide. **A**, Correlation between the width of lines and the width of the resulting facets on eight experimental ochre crayons. Subpanel **a** shows results from crayons with pointed active areas. Subpanel **b** shows results from crayons with linear active areas. The grey bars indicate the width of lines on L13. **B**, Wear facet appearing on the natural surface of an ochre crayon after a single stroke (subpanel **a**) and five strokes (subpanel **b**).



Extended Data Fig. 6 | Three-dimensional rendering of microscopic areas of L13 and silcrete flakes from BBC. Three-dimensional rendering shows flattening of the surface of L13 with the drawing, dissolution of the matrix between quartz grains on the cortex of the BBC silcrete flakes and an unworn appearance of the other surfaces of L13 and the ventral aspect of the BBC silcrete flakes. **A**, Roughness analysis of L13 and MSA silcrete flakes from BBC. Subpanels **a–d** show three-dimensional renderings of a selected area of the surface of L13 with the drawing (**a**), other surfaces of L13 (**b**), knapped (**c**) and cortical (**d**) surfaces of BBC silcrete flakes. Subpanel **e** shows box plots of the variation of roughness variables Sq and Sdr , and a bi-plot correlating these two variables. Notice the high degree of smoothness of the surface of L13 with the drawing relative to the other surfaces. A Kruskal–Wallis multiple comparison test demonstrates that

Sq , Sdr and Spc on the surface with the drawing are significantly lower ($P < 0.01$) than those measured on the remainder of the analysed surfaces of L13. **B**, Areal fractal analysis confirms a clear difference in roughness between the surface with the drawing and other surfaces of L13. This is consistent with the interpretation of the wear on the surface with the drawing as being produced by grinding activities before the drawing occurred. **C**, Analysis of L13 with confocal microscopy. Rectangles indicate the locations of the analysed areas on the surface with the drawing (red) and on the other surfaces (green). Letters distinguish analyses conducted with areal fractal analysis on the surface with the drawing (**a**), a flake scar on the surface with the drawing (**b**), a flake scar on the dorsal surface (**c**) from analyses made on the dorsal (**d**) and ventral surfaces (**e**).



Extended Data Fig. 7 | Ochre and silcrete used in the replication experiments. a, Pieces of ochre used experimentally to produce lines on silcrete flakes. **b,** Silcrete flakes used during the experiments.

A natural variant and engineered mutation in a GPCR promote DEET resistance in *C. elegans*

Emily J. Dennis¹, May Dobosiewicz², Xin Jin^{2,6}, Laura B. Duvall¹, Philip S. Hartman³, Cornelia I. Bargmann^{2,4} & Leslie B. Vosshall^{1,4,5*}

DEET (*N,N*-diethyl-*meta*-toluamide) is a synthetic chemical identified by the US Department of Agriculture in 1946 in a screen for repellents to protect soldiers from mosquito-borne diseases^{1,2}. Since its discovery, DEET has become the world's most widely used arthropod repellent and is effective against invertebrates separated by millions of years of evolution—including biting flies³, honeybees⁴, ticks⁵, and land leeches³. In insects, DEET acts on the olfactory system^{5–12} and requires the olfactory receptor co-receptor *Orco*^{7,9–12}, but exactly how it works remains controversial¹³. Here we show that the nematode *Caenorhabditis elegans* is sensitive to DEET and use this genetically tractable animal to study the mechanism of action of this chemical. We found that DEET is not a volatile repellent, but instead interferes selectively with chemotaxis to a variety of attractant and repellent molecules. In a forward genetic screen for DEET-resistant worms, we identified a gene that encodes a single G protein-coupled receptor, *str-217*, which is expressed in a single pair of chemosensory neurons that are responsive to DEET, called ADL neurons. Mis-expression of *str-217* in another chemosensory neuron conferred responses to DEET. Engineered *str-217* mutants, and a wild isolate of *C. elegans* that carries a *str-217* deletion, are resistant to DEET. We found that DEET can interfere with behaviour by inducing an increase in average pause length during locomotion, and show that this increase in pausing requires both *str-217* and ADL neurons. Finally, we demonstrated that ADL neurons are activated by DEET and that optogenetic activation of ADL neurons increased average pause length. This is consistent with the 'confusant' hypothesis, which proposes that DEET is not a simple repellent but that it instead modulates multiple olfactory pathways to scramble behavioural responses^{10,11}. Our results suggest a consistent motif in the effectiveness of DEET across widely divergent taxa: an effect on multiple chemosensory neurons that disrupts the pairing between odorant stimulus and behavioural response.

We explored whether and how *C. elegans* nematodes respond to DEET during chemotaxis^{14,15} (Fig. 1a) and tested three competing hypotheses for the mechanism of DEET: 'smell-and-repel', which states that the olfactory detection of DEET triggers avoidance^{8,12}; 'masking', which proposes that DEET blocks olfactory pathways that mediate attraction^{6,7}; and 'confusant', which states that DEET modulates multiple olfactory sensory neurons to scramble the perception of an otherwise attractive stimulus^{10,11}.

We first tested the smell-and-repel hypothesis. DEET alone was not repellent when presented to *C. elegans* as a volatile point source (Fig. 1b)—similar to results from *Drosophila melanogaster* flies⁷, *Apis mellifera* bees⁴, and *Aedes aegypti* mosquitoes¹¹, but in contrast to results from *Culex quinquefasciatus* mosquitoes⁸. To investigate whether DEET masks responses to attractive odorants^{6,7} or directly inhibits their volatility⁸, we presented DEET alongside the attractant isoamyl alcohol and found that DEET had no effect on attraction (Fig. 1c). In considering alternative ways to present DEET, we mixed

low concentrations of DEET uniformly into chemotaxis agar (Fig. 1d). DEET-agar reduced chemotaxis to isoamyl alcohol in a dose-dependent manner (Fig. 1e). We tested three additional attractants—butanone, diacetyl and pyrazine—and the volatile repellent 2-nonanone. DEET decreased attraction to butanone and avoidance of 2-nonanone, indicating that DEET affects responses to both positive and negative chemosensory stimuli (Fig. 1f). DEET also affected chemotaxis to diacetyl, but not to pyrazine (Fig. 1f). Notably, both diacetyl and pyrazine are sensed by the AWA sensory neuron. This is similar to *D. melanogaster*, in which DEET can disrupt responses to some odorants—and not others—even in a single chemosensory neuron¹⁰.

In *D. melanogaster* and *A. aegypti*, DEET inhibits attraction to complex natural odour blends^{7,11}. In *C. elegans*, DEET eliminated chemotaxis to the food odour of OP50 *Escherichia coli* bacteria (Fig. 1g). Supplementing bacterial odour with pyrazine, but not isoamyl alcohol, restored chemotaxis (Fig. 1g). To exclude the possibility that pyrazine is able to overcome the effect of DEET because of a higher effective concentration, we carried out dose-response experiments with isoamyl alcohol (Fig. 1h) and pyrazine (Fig. 1i), and found that at all tested concentrations, DEET interfered with isoamyl alcohol chemotaxis but not with pyrazine chemotaxis. DEET chemosensory interference is therefore odour-selective, and affects responses to both attractive and repellent stimuli.

Previous work to identify the genetic basis of sensitivity to DEET using forward genetics has yielded both an X-linked DEET-insensitive *D. melanogaster* mutant¹⁶ and a dominant genetic basis for insensitivity to DEET in mosquitoes¹⁷, but neither study identified the underlying genes. Reverse genetic experiments in *D. melanogaster* and three mosquito species have shown that *orco* is required for sensitivity to DEET^{7,9–12}, but this gene is insect-specific¹⁸—which raises the question of which pathways are used in non-insect invertebrates. We therefore carried out a forward genetic screen for *C. elegans* mutants capable of chemotaxing towards isoamyl alcohol on DEET-agar plates (Fig. 2a). Of five DEET-resistant worms, three produced offspring that chemotaxed towards isoamyl alcohol on DEET-agar plates (Fig. 2b). We used whole-genome sequencing to identify candidate mutations in two strains, and focused on *LBV003*—this maps to *str-217*, which encodes a predicted G protein-coupled receptor (Fig. 2c, d). Over the course of mapping *str-217*, we discovered that a divergent strain of *C. elegans* isolated in Hawaii (CB4856 (Hawaiian)) is naturally resistant to DEET (Fig. 2e) and contains a deletion in *str-217* (*str-217^{HW}*) that leads to a predicted frame shift and early stop codon (Fig. 2c, d, Supplementary Data). This is not a unique phenomenon: 119 out of 247 sequenced strains in the *C. elegans* Natural Diversity Resource¹⁹ contained predicted changes in *STR-217*, compared to the N2 wild type (Supplementary Data). We next tested three near-isogenic lines with a single, homozygous genomic segment of Hawaiian chromosome V introgressed into a wild-type background²⁰ (Fig. 2e). Only the ewIR74 line contains *str-217^{HW}* and was resistant to DEET (Fig. 2e). Expressing a rescue-reporter strain (Fig. 2f) in *LBV003* (Fig. 2g) or *ewIR74* (Fig. 2h) rendered both strains fully sensitive to DEET (Fig. 2g, h).

¹Laboratory of Neurogenetics and Behaviour, The Rockefeller University, New York, NY, USA. ²Lulu and Anthony Wang Laboratory of Neural Circuits and Behaviour, The Rockefeller University, New York, NY, USA. ³Department of Biology, Texas Christian University, Fort Worth, TX, USA. ⁴Kavli Neural Systems Institute, New York, NY, USA. ⁵Howard Hughes Medical Institute, New York, NY, USA. ⁶Present address: Society of Fellows, Harvard University, Cambridge, MA, USA. *e-mail: Leslie.Vosshall@rockefeller.edu

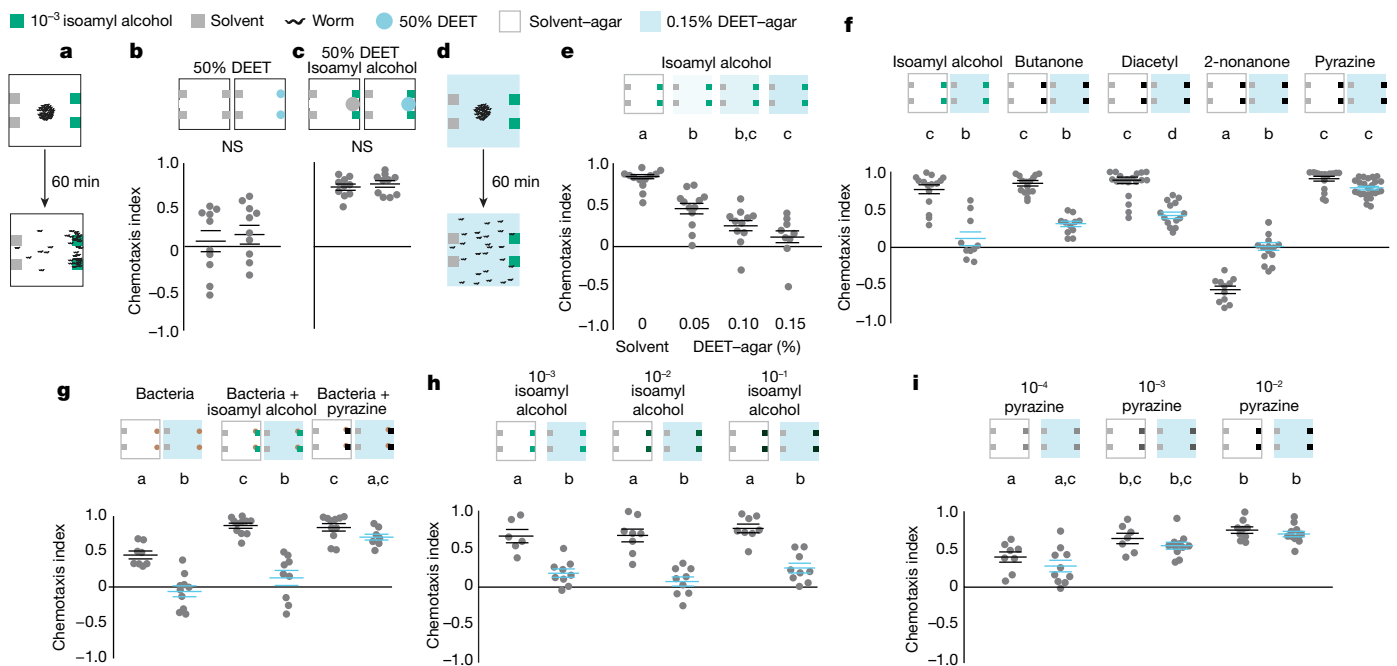


Fig. 1 | DEET interferes with chemotaxis in wild-type *C. elegans*.
a, Schematic of chemotaxis assay. **b, c**, Chemotaxis with point-source stimuli of DEET without (**b**) or with (**c**) isoamyl alcohol. **d**, Chemotaxis assay on DEET-agar plates. **e**, Chemotaxis to isoamyl alcohol with different concentrations of DEET in the agar. **f**, Chemotaxis to indicated odors. **g**, Chemotaxis to bacterial food source with indicated odors. **h, i**, Chemotaxis dose-response to isoamyl alcohol (**h**) and pyrazine (**i**). In **b, c**, and **e–i**, each dot represents a chemotaxis index of a single population

assay. Schematics indicate experimental conditions in the corresponding column in the plot, according to the key provided at the top of the figure (except where indicated otherwise, the key applies to all figures). In a given panel, data labelled with different letters are significantly different from each other; mean \pm s.e.m. (horizontal blue or black lines on plots), $P < 0.05$, two-sided t -test (**b, c**) or one-way (**e**) or two-way (**f–i**) ANOVA and Tukey's post hoc test. NS, not significant. See Supplementary Data for sample sizes and details of statistical analyses.

In insects, DEET interacts with chemosensory neurons^{7,9–12}. We used calcium imaging to investigate whether DEET modulates the primary sensory neurons (AWC neurons) required for the detection of isoamyl alcohol²¹. AWC neurons responded to the addition of DEET with a rapid increase in calcium that decreased to baseline over the course of 11 min of chronic DEET stimulation (Fig. 3a). In the presence of DEET, the responses of AWC neurons to isoamyl alcohol

decreased in magnitude, but there were no significant differences in AWC-neuron activity between wild type and *str-217*^{−/−}, a predicted null mutant produced by CRISPR–Cas9 genome editing (Fig. 3a–c). The polymodal nociceptive neurons, named ASH, also responded to DEET (Fig. 3d, e), but worms that lack ASH neurons are fully DEET-sensitive (Fig. 3f). This suggests that DEET resistance requires additional neurons.

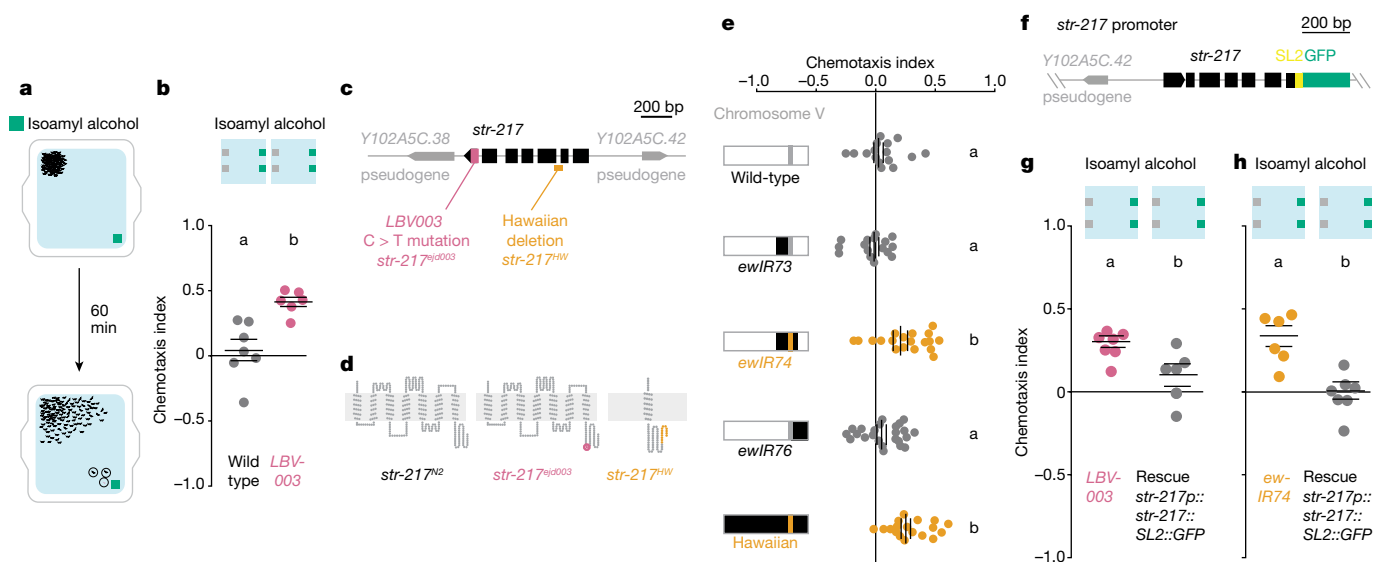


Fig. 2 | *str-217* mutants are resistant to DEET. **a**, Schematic of forward genetic screen with DEET-resistant mutants circled. **b**, Chemotaxis of indicated strains. **c**, Genomic locus of *str-217*. **d**, STR-217 protein snake plots in indicated strains. **e**, Schematic of chromosome V (left), and chemotaxis (right) in indicated strains. **f**, *str-217* rescue-reporter construct. **g, h**, Chemotaxis of indicated strains. In **b, e, g, h**, each dot

represents a chemotaxis index of a single population assay. In a given panel, data labelled with different letters are significantly different from each other; mean \pm s.e.m., $P < 0.05$, ANOVA and Tukey's post hoc test (**e**) and two-sided t -test (**b, g, h**). See Supplementary Data for sample sizes and details of statistical analyses.

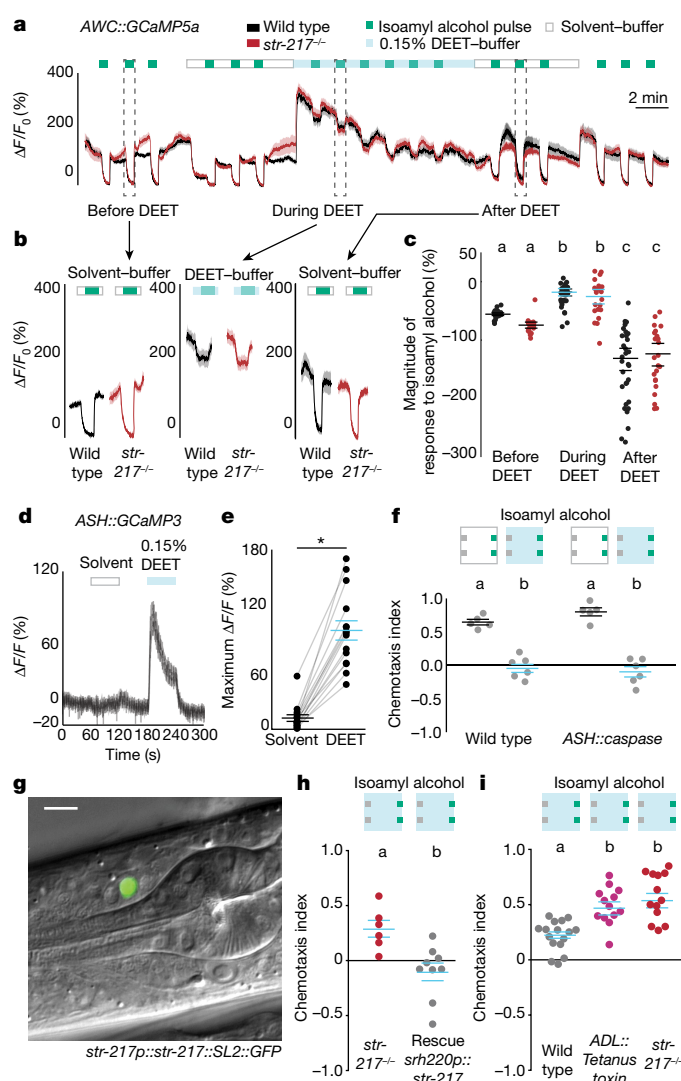


Fig. 3 | Several *C. elegans* neurons respond to DEET but ADL neurons are required for sensitivity to DEET. **a**, Average traces of GCaMP activity in AWC^{ON} cells of indicated genotype over a 36-min experiment. **b**, A subset of traces from **a**, re-plotted separately. **c**, Response magnitudes of data shown in **b**. **d**, GCaMP activity in ASH neurons to indicated stimuli. **e**, Quantification of **d**. **f**, Chemotaxis of indicated strains to isoamyl alcohol. **g**, *str-217* rescue-reporter GFP expression in a single ADL neuron. Scale bar, 10 μ m. **h**, **i**, Isoamyl alcohol chemotaxis of indicated strains. In **a**, **b**, **d**, each trace represents mean (dark line) \pm s.e.m. (lighter area) GCaMP response of all worms of each genotype. In **c**, **e**, each dot represents the response of a single worm. In **f**, **h**, **i**, each dot represents a chemotaxis index of a single population assay. In a given panel, data labelled with different letters are significantly different from each other; mean \pm s.e.m., $P < 0.05$, two-way (**b**, **e**) or one-way (**h**) ANOVA with Tukey's post hoc test, or paired *t*-test (**d**, **g**). In **e**, $*P < 0.05$. See Supplementary Data for sample sizes and details of statistical analyses.

To identify such neurons, we determined where *str-217* is expressed by examining our *str-217* rescue-reporter strain; we found GFP expression in a single pair of ADL chemosensory neurons (Fig. 3g). This was unexpected because ADL neurons are not required for chemotaxis to isoamyl alcohol²², which suggests an indirect role in DEET chemosensory interference. To confirm that *str-217* expression in ADL neurons is required for DEET sensitivity, we expressed *str-217* in ADL neurons in *str-217*^{-/-} mutants using a different ADL promoter, *srh-220p*. Expressing *str-217* in ADL neurons rendered this mutant sensitive to DEET (Fig. 3h). To investigate whether ADL neurons are required for DEET sensitivity, we inhibited chemical synaptic transmission by expressing the tetanus toxin light chain in ADL neurons^{23,24}.

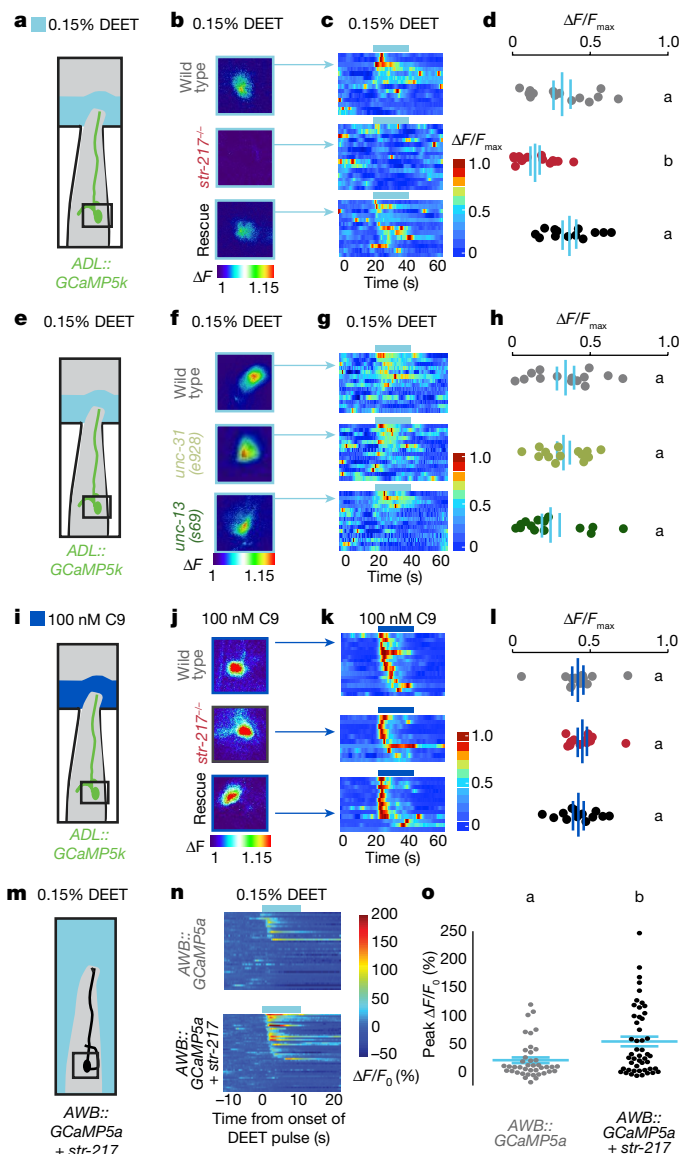


Fig. 4 | *str-217* is necessary for ADL responses to DEET and can confer DEET sensitivity to AWB neurons. **a**, **e**, **i**, **m**, Schematics of calcium imaging. **b**, **f**, **j**, Pseudo-coloured images of ADL neurons responding to 0.15% DEET (**b**, **f**) or 100 nM C9 (**j**) in worms of indicated genotype (fold increase in mean fluorescence 20 s during first stimulus pulse/mean of 20 s before the stimulus pulse). Arrows point to corresponding worm in subsequent panels. **c**, **g**, **k**, Heat maps of GCaMP activity in response to DEET (**c**, **g**) or C9 (**k**). Each row represents activity in one worm. **d**, **h**, **l**, Mean normalized responses of the data in **c**, **g**, **k** during entire pulse of DEET (**d**, **h**) or C9 (**l**). **n**, Heat maps of AWB-neuron calcium imaging response to 0.15% DEET; each row represents imaging from one worm cropped to show the 10-s before, during, and after the DEET pulse. **o**, Quantification of **n**. In **d**, **h**, **l**, **o**, each dot represents a single neuron in a single worm. In a given panel, data labelled with different letters are significantly different from each other; mean \pm s.e.m. $P < 0.05$, one-way ANOVA and Tukey's post hoc test (**d**, **h**, **l**) or two-sided *t*-test (**o**). The rescue construct in **c** is the same as that shown in Fig. 3g, h. See Supplementary Data for sample sizes and details of statistical analyses.

These worms showed similar DEET resistance to that of *str-217* mutants (Fig. 3i), which demonstrates that ADL neurons contribute to DEET sensitivity.

Because both *str-217* and ADL neurons contribute to DEET sensitivity, we used calcium imaging to investigate whether ADL neurons respond to DEET and whether this requires *str-217* (Fig. 4a). Wild type and *str-217*^{-/-} mutants that carry a rescue plasmid—but not *str-217*^{-/-} mutants—showed calcium responses to DEET in ADL neurons,

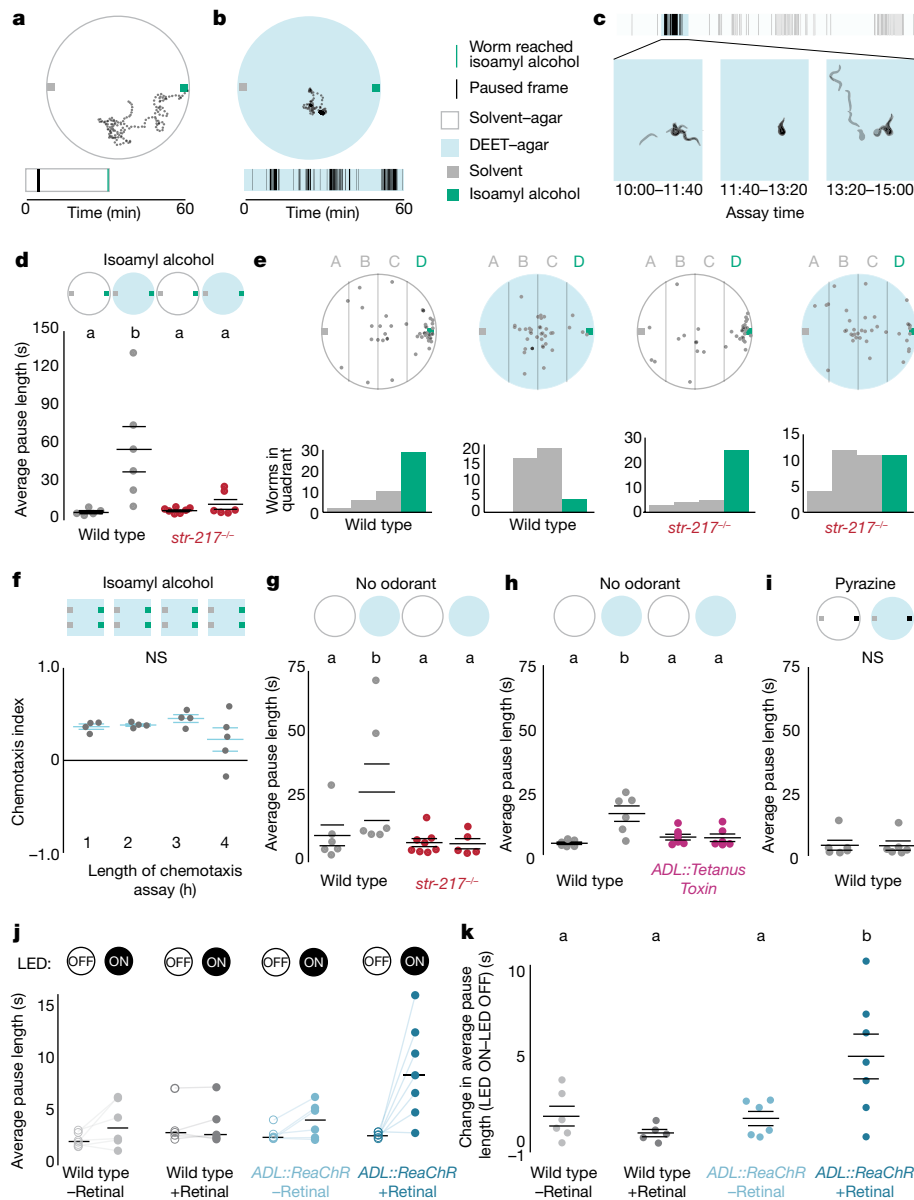


Fig. 5 | DEET increases average pause length by activating *str-217* and ADL neurons. **a, b**, Top, example trajectories of a single wild-type worm chemotaxing to isoamyl alcohol on solvent-agar (a) or DEET-agar (b). Each dot depicts the x, y position every 10 s. Bottom, raster plots of paused frames for the worm shown above. **c**, Example pauses from **b**. Images were extracted and converted to silhouettes every 6 s and superimposed. **d**, Average pause length of indicated stimuli and genotypes. **e**, Top, location of worms in **d** after 60 min. Bottom, histograms of worm locations. **f**, Wild-type chemotaxis on DEET at indicated times.

albeit with some variability in the response (Fig. 4a–d). To exclude the possibility that DEET activates ADL neurons indirectly, we measured calcium responses in ADL neurons in *unc-13* and *unc-31* mutants, which are deficient in synaptic vesicle fusion²⁵ and dense-core vesicle fusion²⁶, respectively. ADL-neuron responses to DEET were normal in both strains (Fig. 4e–h). A known agonist of ADL neurons—the pheromone C9²³—activated ADL neurons in wild-type, *str-217*^{−/−} mutant and rescued worms (Fig. 4i–l), which suggests that the lack of DEET responsiveness in *str-217* is selective.

We then investigated whether *str-217* is sufficient to confer DEET responses in vitro in HEK293T cells or in vivo in AWB olfactory sensory neurons, which respond only weakly to DEET. Although DEET did not activate HEK293T cells that express *str-217* (Supplementary Data), in vivo mis-expression of *str-217* in AWB neurons significantly increased the sensitivity to DEET of these sensory neurons (Fig. 4m–o).

g, i, Average pause length of indicated stimuli and genotypes. **j**, Average pause length of indicated genotypes and LED status (off or on). Lines connect experimental pairs. **k**, Difference in average pause length for each experimental pair in **j**. In **d, f–k**, each dot represents a single experimental plate. In a given panel, data labelled with different letters are significantly different from each other; mean \pm s.e.m. $P < 0.05$, one-way (f) or two-way (d, g–k) ANOVA and Tukey's post hoc test. See Supplementary Data for sample sizes and details of statistical analyses.

This suggests that *str-217* either encodes a DEET receptor or cooperates with other proteins to increase sensitivity to DEET.

We next tracked the position and posture of individual worms on DEET-agar or solvent-agar plates (Fig. 5a–c). Wild-type worms, but not *str-217*^{−/−} mutants (Fig. 5d), showed marked increases in average pause length on DEET-agar. Although *str-217*^{−/−} mutants are resistant to DEET compared to wild type, their chemotaxis is not fully resistant to DEET (Fig. 5h) and many *str-217*^{−/−} worms never reached the odorant source (Fig. 5e). Additionally, prolonging the assays did not significantly increase performance in wild-type worms (Fig. 5f). We suspect that DEET has additional effects on chemotaxis in addition to increasing pause duration.

To test whether the increase in pausing on DEET occurs only during chemotaxis, we tracked wild-type, *str-217*^{−/−} mutant (Fig. 5g), and *ADL::Tetanus toxin* (Fig. 5h) worms on DEET-agar and solvent-agar

plates without odorants. *ADL::Tetanus toxin* worms express an inhibitor of evoked synaptic vesicle release in ADL neurons. Of these, only wild-type worms had an average pause length that was longer on DEET–agar than on solvent–agar (Fig. 5g, h), and this was not seen during pyrazine chemotaxis (Fig. 5i).

To test whether ADL activity alone is sufficient to increase average pause length, we optogenetically activated the light-sensitive ion channel ReaChR²⁷ in ADL neurons in wild-type worms and observed an increase in average pause length (Fig. 5j, k). We conclude that ADL neurons mediate the increase in average pause length seen on DEET–agar, and speculate that the increase in long pauses is one mechanism by which DEET interferes with chemotaxis.

Here we add the nematode *C. elegans* to known DEET-sensitive animals and uncover a neuronal mechanism for DEET-induced behaviour. This work opens up *C. elegans* as a system to test new repellents in vivo and to discover additional genes and neurons that respond to DEET. The molecular mechanism by which the *str-217* mutation renders ADL neurons insensitive to DEET and worms resistant to DEET remains to be understood. *str-217* encodes an orphan G protein-coupled receptor that is evolutionarily unrelated to DEET-sensitive insect odorant receptors. Our results are consistent with the hypothesis that *str-217* is a DEET receptor or the hypothesis that it interacts with additional proteins to make ADL neurons sensitive to DEET. Pyrazine chemotaxis is not disrupted by DEET, consistent with our model in which DEET is not a simple repellent but is instead a modulator of behaviour that interferes with chemotaxis to some—but not all—odorants. The *str-217*-dependent mechanism of action of DEET in nematodes is reminiscent of the ‘confusant’ hypothesis in insects, in which DEET alters responses of individual olfactory sensory neurons to attractive odorants^{7,10}, thereby interfering with behavioural attraction. In *C. elegans*, DEET inhibits attraction to some odorants by activating neurons that induce competing behaviours, such as pausing. We speculate that the promiscuity of DEET in interacting with multiple molecules and chemosensory neurons across vast evolutionary scales is the key to the broad effectiveness of this chemical.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0546-8>.

Received: 4 October 2017; Accepted: 27 July 2018;

Published online 26 September 2018.

- Travis, B. V. et al. The more effective mosquito repellents tested at the Orlando, Fla., laboratory, 1942–47. *J. Econ. Entomol.* **42**, 686–694 (1949).
- McCabe, E. T., Barthel, W. F., Gertler, S. I. & Hall, S. A. Insect Repellents. III. *N, N*-diethylamides. *J. Org. Chem.* **19**, 493–498 (1954).
- Tawatsin, A. et al. Field evaluation of DEET, Repel Care, and three plant based essential oil repellents against mosquitoes, black flies (Diptera: Simuliidae) and land leeches (Arhynchobdellida: Haemadipsidae) in Thailand. *J. Am. Mosq. Control Assoc.* **22**, 306–313 (2006).
- Abramson, C. I. et al. Proboscis conditioning experiments with honeybees, *Apis mellifera caucasica*, with butyric acid and DEET mixture as conditioned and unconditioned stimuli. *J. Insect Sci.* **10**, 122 (2010).
- Carroll, J. F., Klun, J. A. & Deboun, M. Repellency of DEET and SS220 applied to skin involves olfactory sensing by two species of ticks. *Med. Vet. Entomol.* **19**, 101–106 (2005).
- Dogan, E. B., Ayres, J. W. & Rossignol, P. A. Behavioural mode of action of DEET: inhibition of lactic acid attraction. *Med. Vet. Entomol.* **13**, 97–100 (1999).
- Ditzen, M., Pellegrino, M. & Vosshall, L. B. Insect odorant receptors are molecular targets of the insect repellent DEET. *Science* **319**, 1838–1842 (2008).
- Syed, Z. & Leal, W. S. Mosquitoes smell and avoid the insect repellent DEET. *Proc. Natl Acad. Sci. USA* **105**, 13598–13603 (2008).
- Liu, C. et al. Distinct olfactory signaling mechanisms in the malaria vector mosquito *Anopheles gambiae*. *PLoS Biol.* **8**, e1000467 (2010).

- Pellegrino, M., Steinbach, N., Stensmyr, M. C., Hansson, B. S. & Vosshall, L. B. A natural polymorphism alters odour and DEET sensitivity in an insect odorant receptor. *Nature* **478**, 511–514 (2011).
- DeGennaro, M. et al. *orco* mutant mosquitoes lose strong preference for humans and are not repelled by volatile DEET. *Nature* **498**, 487–491 (2013).
- Xu, P., Choo, Y. M., De La Rosa, A. & Leal, W. S. Mosquito odorant receptor for DEET and methyl jasmonate. *Proc. Natl Acad. Sci. USA* **111**, 16592–16597 (2014).
- DeGennaro, M. The mysterious multi-modal repellency of DEET. *Fly* **9**, 45–51 (2015).
- Bargmann, C. I. & Horvitz, H. R. Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in *C. elegans*. *Neuron* **7**, 729–742 (1991).
- Cho, C. E., Brueggemann, C., L’Etoile, N. D. & Bargmann, C. I. Parallel encoding of sensory history and behavioral preference during *Caenorhabditis elegans* olfactory learning. *eLife* **5**, e14000 (2016).
- Reeder, N. L., Ganz, P. J., Carlson, J. R. & Saunders, C. W. Isolation of a DEET-insensitive mutant of *Drosophila melanogaster* (Diptera: Drosophilidae). *J. Econ. Entomol.* **94**, 1584–1588 (2001).
- Stanczyk, N. M., Brookfield, J. F., Ignell, R., Logan, J. G. & Field, L. M. Behavioral insensitivity to DEET in *Aedes aegypti* is a genetically determined trait residing in changes in sensillum function. *Proc. Natl Acad. Sci. USA* **107**, 8575–8580 (2010).
- Brand, P. et al. The origin of the odorant receptor gene family in insects. *eLife* **7**, e38340 (2018).
- Cook, D. E., Zdrzaljevic, S., Roberts, J. P. & Andersen, E. C. CeNDR, the *Caenorhabditis elegans* Natural Diversity Resource. *Nucleic Acids Res.* **45**, D650–D657 (2017).
- Doroszkow, A., Snoek, L. B., Fradin, E., Riksen, J. & Kammenga, J. A genome-wide library of CB4856/N2 introgression lines of *Caenorhabditis elegans*. *Nucleic Acids Res.* **37**, e110 (2009).
- Bargmann, C. I., Hartwig, E. & Horvitz, H. R. Odorant-selective genes and neurons mediate olfaction in *C. elegans*. *Cell* **74**, 515–527 (1993).
- Zaslaver, A. et al. Hierarchical sparse coding in the sensory system of *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **112**, 1185–1189 (2015).
- Jang, H. et al. Neuromodulatory state and sex specify alternative behaviors through antagonistic synaptic pathways in *C. elegans*. *Neuron* **75**, 585–592 (2012).
- Schiavo, G. et al. Tetanus toxin is a zinc protein and its inhibition of neurotransmitter release and protease activity depend on zinc. *EMBO J.* **11**, 3577–3583 (1992).
- Richmond, J. E., Davis, W. S. & Jorgensen, E. M. UNC-13 is required for synaptic vesicle fusion in *C. elegans*. *Nat. Neurosci.* **2**, 959–964 (1999).
- Speese, S. et al. UNC-31 (CAPS) is required for dense-core vesicle but not synaptic vesicle exocytosis in *Caenorhabditis elegans*. *J. Neurosci.* **27**, 6150–6162 (2007).
- Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).

Acknowledgements We thank M. Crickmore, K. J. Lee, A. Singhvi, N. Yapici and members of the Vosshall Laboratory for comments on the manuscript, and for experimental assistance and advice: S. Shaham and W. Wang for assistance with chemical mutagenesis; H. Jang for assistance with chemotaxis behaviour and imaging; A. Lopez-Cruz and E. Scheer for assistance with tracking; S. Levy and E. Scheer for plasmids and strains; and A. Nguyen for early analysis of mutants (with P.S.H.). This work was conducted with support from NIH (to E.J.D., F31 DC014222) and the CGC (P40 OD010440), which provided selected strains. L.B.V. is an investigator of the Howard Hughes Medical Institute.

Reviewer information Nature thanks E. Poivet and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.J.D. and L.B.V. conceived the project, wrote the manuscript and produced the figures: E.J.D. performed the experiments and analyses shown in all figures, except for Figs. 3a–e, 4m–o (performed by M.D.) and Fig. 3g (performed by X.J.). L.B.V. performed HEK293T expression. C.I.B. provided guidance, experimental design advice and data interpretation. P.S.H. made the original observation that DEET interferes with chemotaxis, and performed initial genetic screens.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0546-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.B.V.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size except the experiments shown in Fig. 2e and the ADL-imaging experiments in Fig. 4, for which a power calculation was performed based on pilot data performed under similar but not identical conditions. Animals were assigned randomly to different experimental conditions and locations of behavioural experiments were pseudo-randomized each day to avoid positioning biases. Inclusion and exclusion criteria were pre-established for all experiments, and plate positions were pseudo-randomized in behaviour experiments. The investigators were not blinded to allocation during experiments and outcome assessment except for population behavioural experiments, in which a subset of the data gathered was blinded at either the allocation or scoring phase. For all tracking experiments, the experimenter was blinded to genotype during any manual linking and curation.

Nematode culture and strains. *C. elegans* strains were maintained at room temperature (22–24 °C) on nematode growth medium (NGM) plates (51.3 mM NaCl, 1.7% agar, 0.25% peptone, 1 mM CaCl₂, 12.9 μM cholesterol, 1 mM MgSO₄, 25 mM KPO₄, pH 6) seeded with *E. coli* (OP50 strain) as a food source^{28,29}. Bristol N2 was used as the wild-type strain. The CB4856 (Hawaiian) strain, containing WBVar02076179 (*str-217^{HW}*) (<http://www.wormbase.org/db/get?name=WBVar02076179;class=variation>) and Hawaiian recombinant inbred strains for chromosome V were previously generated²⁰. Generation of extra-chromosomal array transgenes was carried out using standard procedures³⁰, and included the transgene injected at 50 ng/μl, the fluorescent co-injection marker *Pelt-2::GFP* at 5 ng/μl (with the exception of LBV004 and LBV009, which did not include a co-injection marker), and an empty vector for a total DNA concentration of 100 ng/μl. CRISPR–Cas9-mediated mutagenesis of *str-217* was performed as described, using *rol-6* as a co-CRISPR marker³¹. The resulting *str-217* mutant strain (LBV004 *str-217(ejd001)*) results in a predicted frame-shift in the first exon (indel: insertion (AAAAAAA), deletion (CTGCTCCA), final sequence GCGTCGAAAAAAATTCAG; insertion is underlined). The *str-217* rescue construct (*Pstr-217::str-217::SL2::GFP*) used a 1,112-nucleotide-length fragment 56 nucleotides upstream 5' of the translation start of *str-217*.

Microscopy and image analysis. L2-adult stage hermaphrodites were mounted on 1% agarose pads with 10 mM sodium azide (CID 6331859, Sigma-Aldrich, S2002) in M9 solution (22 mM KH₂PO₄, 42 mM Na₂HPO₄, 85.6 mM NaCl, 1 μM MgSO₄, pH 6). Images were acquired with an Axio Observer Z1 LSM 780 with Apotome a 63× objective (Zeiss), and were processed using ImageJ.

Chemotaxis assays. Chemotaxis was tested as described¹⁵, on square plates containing 10 ml of chemotaxis agar (1.6% agar in chemotaxis buffer: 5 mM phosphate buffer pH 6.0, 1 mM CaCl₂, 1 mM MgSO₄)³². Additions of either ethanol (solvent–agar) or 50% DEET (CID: 4284, Sigma-Aldrich, D100951) in ethanol (DEET–agar) were added immediately before pouring, after agar cooled to <44 °C. A total volume of 300 μl ethanol or DEET in ethanol was added to each 100 ml of agar mixture for all experiments except Figs. 1b, c, 5j, k. Plates were poured on the day of each experiment, and dried with lids off for 4 h before the start of the assay. One microlitre 1 M sodium azide was added to two spots on either side of the plate right before beginning the experiment to immobilize worms that reached the odorant or ethanol sources. Three days before all chemotaxis experiments, 4–6 L4 worms were transferred onto NGM plates seeded with *E. coli* (OP50 strain). The offspring of these 4–6 worms were then washed off the plates and washed twice with S-Basal buffer (1 mM NaCl, 5.74 mM K₂HPO₄, 7.35 mM KH₂PO₄, 5 μg/ml cholesterol at pH 6–6.2)²⁸ to remove younger worms, and once with chemotaxis buffer. Immediately before the start of the experiment, two 1-μl drops of odorant diluted in ethanol, or ethanol solvent control, were spotted on each side of the plate on top of the sodium azide spots. Between 100 and 300 worms were then placed into the centre of the plate in a small bubble of liquid. The excess liquid surrounding the worms was then removed using a Kimwipe. Odorants diluted in ethanol were used in this study at these concentrations unless otherwise noted: 1:1,000 isoamyl alcohol (CID: 31260, Sigma-Aldrich, W205702), 1:1,000 butanone (CID: 6569, Sigma-Aldrich, 360473), 10 mg/μl pyrazine (CID: 9261, Sigma-Aldrich, W401501), 1:10 2-nonanone (CID: 13187, Sigma-Aldrich, W2787513). For bacterial chemotaxis assays, 20 μl of either LB medium, or OP50 bacterial suspension grown overnight and diluted in LB medium to an optical density (OD) at 600 nm of 1.0, was applied instead of or in addition to odorants. Assays were carried out for 60–90 min at room temperature (22–24 °C) between 13:00 and 20:00 Eastern Standard Time with the exception of those in Fig. 5f, which were quantified after either 55–65 min (1 h), 115–125 min (2 h), 175–185 min (3 h) or 235–245 min (4 h). Plates were scored as soon as possible, either immediately or—if a large number of plates was being scored on the same day—plates were moved to 4 °C to immobilize worms until they could be scored. The assay was quantified by counting worms that had left the origin in the centre of the plate, moving to either side of the plate (denoted '#Odorant' or '#Control') or just above or below the origin (denoted '#Other'), and calculating a chemotaxis index as (#Odorant – #Control)/(#Odorant + #Control + #Other). A trial was discarded

if fewer than 50 worms or more than 250 worms contributed to the chemotaxis index and participated in the assay.

Mutant screen. About 100 wild-type (Bristol N2) L4 worms were mutagenized in M9 solution with 50 mM ethyl methanesulfonate (EMS) (CID: 6113, Sigma-Aldrich, M0880) for 4 h with rotation at room temperature. Mutagenized worms were picked to separate 9-cm NGM agar plates seeded with *E. coli* (OP50 strain) and cultivated at 20 °C. About 5,000 F2 worms were screened for DEET resistance on 20.3-cm casserole dishes (ASIN B000LNS4NQ, model number 81932OBL11). Five worms across three assays were more than ~2 cm closer to the odorant source than the rest of the worms on the plate and were defined as DEET-resistant. This phenotype was heritable in three strains, and each strain was backcrossed to OS1917 for four generations. Whole-genome sequencing³³ was used to map the mutations to regions containing transversions presumably introduced by the EMS, parental alleles of the N2 strain used for mutagenesis and missing alleles of the wild-type strain OS1917 used for backcrossing^{34,35}. LBV003 mapped to a 5-Mb region on chromosome V, which was further mapped to *str-217*. LBV002 mapped to a 6.8-Mb region on chromosome V, which was further narrowed down to a likely candidate gene, *nstp-3(ejd002)*. In LBV002, *nstp-3(ejd002)* contains a T > G transversion of the 141st nucleotide in the coding sequence, which is predicted to produce a Phe48Val substitution in this predicted sugar:proton symporter. We were unable to map the DEET-resistant mutation(s) in LBV001.

***str-217* heterologous expression in mammalian tissue culture cells.** HEK293T cells were maintained using standard protocols in a Thermo Scientific FORMA Series II water-jacketed CO₂ incubator. HEK293T cells were obtained directly from Invitrogen and were not tested for mycoplasma contamination. Cells were transiently transfected with 1 μg each of pME18S plasmid expressing *GCaMP6s*, *Gqα15* and *str-217* using Lipofectamine 2000 (CID: 100984821, Invitrogen, 1168019). Control cells excluded *str-217*, but were transfected with the other two plasmids. Transfected cells were seeded into 384-well plates at a density of 2 × 10⁶ cells/ml, and incubated overnight in FluoroBrite DMEM medium (Thermo Fisher Scientific) supplemented with fetal bovine serum (Invitrogen, 10082139) at 37 °C and 5% CO₂. Cells were imaged in reading buffer (Hanks's Balanced Salt Solution (GIBCO) + 20 mM HEPES (Sigma-Aldrich)) using GFP-channel fluorescence of a Hamamatsu FDSS-6000 kinetic plate reader at The Rockefeller University High-Throughput Screening Resource Centre. DEET was prepared at 3 × final concentration in reading buffer in a 384-well plate (Greiner Bio-one) from a 46% (2 M) stock solution in DMSO (Sigma-Aldrich). Plates were imaged every 1 s for 5 min. Ten microlitres of DEET solution in reading buffer or vehicle (reading buffer + DMSO) was added to each well containing cells in 20 μl of medium, after 30 s of baseline fluorescence recording. The final concentration of vehicle DMSO was matched to the DEET additions, with a maximum DMSO concentration of 7.8%. Fluorescence was normalized to baseline, and responses were calculated as maximum ratio (maximum fluorescence level/baseline fluorescence level) (Supplementary Data).

Calcium imaging in ADL neurons. Calcium imaging and data analysis were performed as described³⁶, using single young adult hermaphrodites immobilized in a custom-fabricated 3 × 3 × 3 mm³ polydimethylsiloxane (PDMS) imaging chip. *GCaMP5k* was expressed in ADL neurons under control of the *sre-1* promoter²³ and was crossed into *str-217^{-/-}* and the *str-217^{-/-}* rescue strain. Imaging of *unc-13* and *unc-31* mutant strains was performed by crossing *ADL::GCaMP5k*-expressing worms to the *unc-13* and *unc-31* strains and selecting for fluorescent, uncoordinated worms. Worms were acclimatized to the imaging room overnight on *E. coli* (OP50 strain)-seeded plates. All stimuli were prepared the day of each experiment, and were diluted in ethanol to 1,000× the desired concentration before being further diluted 1:1,000 in S-Basal buffer. Young adult worms were paralysed briefly in (–)-tetramisole hydrochloride (CID: 27944, Sigma-Aldrich, L9756) at 1 mM for 2–5 min before transfer into the chip to paralyse body wall muscles to keep worms stationary during imaging. All worms were pre-exposed to light (470 ± 40 nm) for 100 s before recording to attenuate the light response of ADL neurons³⁷. Experiments consisted of the following stimulation protocol: 20 s S-Basal buffer, followed by 3 repetitions of 20 s DEET (0.15% DEET and 0.15% ethanol in S-Basal) and then 20 s S-basal buffer. ADL-neuron responses desensitize rapidly³⁷, so only the first of the three DEET pulses was analysed.

All GCaMP signals were recorded with Metamorph Software (Molecular Devices) and an iXon3 DU-897 EMCCD camera (Andor) at 10 frames/s using a 40× objective on an upright Zeiss Axioskop 2 microscope. Custom ImageJ scripts¹⁵ were used to track cells and quantify fluorescence. In Fig. 4b, f, j, all frames in 20 s before the DEET pulse were averaged and divided by the average of the frames during the 20-s DEET or C9 pulse to calculate ΔF. In Fig. 4c, g, k, traces were bleach-corrected using a custom MATLAB script and then the 5% of frames with the lowest values were averaged to create F₀. ΔF/F₀ was calculated by (F – F₀)/F₀ and then divided by the maximum value to obtain³⁸ ΔF/F_{max}. The average value during the stimulus was calculated for each worm and plotted as a single dot in Fig. 4d, h, l. The heat-map traces in Fig. 4c, g, k were also smoothed

by 5 frames, such that each data point (n) is the running average of $n - 2$, $n - 1$, n , $n + 1$ and $n + 2$.

Calcium imaging in AWC, ASH and AWB neurons. Calcium imaging of freely moving worms and subsequent data analysis were performed as described³⁸, using a 3-mm² microfluidic PDMS device with two arenas that enabled simultaneous imaging of two genotypes with approximately 10 worms each. For imaging in AWC neurons, we used an integrated line (CX17256) that expresses *GCaMP5a* in AWC^{ON} neurons under control of the *str-2* promoter, crossed into *str-217*^{-/-} worms. Adult hermaphrodites were first paralysed for 80–100 min in 1 mM (–)–tetramisole hydrochloride and then transferred to the arenas in S-Basal buffer. The stimulus protocol was as follows: in S-Basal, three pulses of 60 s in buffer and 30 s in isoamyl alcohol, followed by 120 s in buffer. Next, the worms were switched to S-Basal with 0.15% ethanol (solvent buffer) and three pulses of 60 s in buffer and 30 s in isoamyl alcohol in solvent buffer, followed by 120 s in solvent buffer before a switch to S-Basal with 0.15% ethanol and 0.15% DEET (DEET buffer). In DEET buffer, worms were given 6 pulses of 60 s in DEET buffer and then 30 s in isoamyl alcohol in DEET buffer, followed by 120 s in DEET buffer before switching to solvent buffer. In solvent buffer, the worms received three pulses of 60 s in buffer and 30 s in isoamyl alcohol in solvent buffer, followed by 120 s in solvent buffer before a switch to S-Basal. In S-Basal, the worms received three pulses of 60 s in buffer and 30 s isoamyl alcohol, followed by 60 s in buffer.

Each experiment was repeated 3–4 times over 2–3 days and pooled by strain for analysis (wild type: 31 worms, 4 experiments, 3 days; *str-217*^{-/-}: 23 worms, 3 experiments, 2 days). Images were acquired at 10 frames/s at 5× magnification (Hamamatsu Orca Flash 4 sCMOS), with 10-ms pulsed illumination every 100 ms (Sola, Lumencor; 470/440-nm excitation). Fluorescence levels were analysed using a custom ImageJ script that integrates and subtracts the background fluorescence levels of the AWC neuron cell body (6 × 6 pixel region of interest). Traces were normalized by subtracting and then dividing by the baseline fluorescence, defined as the average fluorescence of the last 2 s of the first 3 isoamyl alcohol pulses. The traces in Fig. 3a, b were also smoothed by 5 frames, such that each data point (n) is the running average of $n - 2$, $n - 1$, n , $n + 1$ and $n + 2$. The response magnitudes in Fig. 3c were calculated by taking the mean of the last 2 s of an isoamyl alcohol pulse, subtracting the mean of the 2 s before the isoamyl alcohol pulse (F_0), and dividing by this F_0 . The response magnitudes were calculated for the 5th (0.15% ethanol in S-Basal buffer) and 8th (0.15% DEET and 0.15% ethanol in S-Basal buffer) isoamyl alcohol pulses.

Calcium imaging in ASH neurons was performed similarly, with the following exceptions. For imaging in ASH neurons, we used a strain (CX10979) expressing *GCaMP3* in ASH neurons under control of the *sra-6* promoter. The stimulus protocol used was as follows: 60 s in S-Basal, 60 s in 0.15% ethanol in S-Basal buffer, 60 s in S-Basal, 60 s in 0.15% DEET in S-Basal and finally 60 s in S-Basal buffer. Each experiment was repeated over 2 days and pooled for analysis (wild type: 15 worms in 2 experiments on 2 different days).

Calcium imaging in AWB neurons was performed similarly to imaging in AWC neurons, with the following exceptions. For imaging in AWB neurons, we used a control strain (CX17428) expressing *GCaMP5a* in AWB neurons under the *str-1* promoter and a test strain (CX17660) expressing *GCaMP5a* under the *str-1* promoter, as well as expressing *str-217* in AWB neurons under the *str-1* promoter. Adult hermaphrodites were first similarly paralysed in 1 mM (–)–tetramisole hydrochloride for 80–100 min, but the first 65–75 min was in S-Basal buffer and the last 15 min was 1 mM (–)–tetramisole hydrochloride in ethanol–buffer. The stimulus protocol used was as follows: 60 s in 0.15% ethanol in S-Basal buffer, 10 s in 0.15% DEET and 0.15% ethanol in S-Basal buffer, and 60 s in 0.15% ethanol in S-Basal buffer. A 4 × 4-pixel region of interest was used during tracking of the neurons. Baseline fluorescence was defined as the median fluorescence of the 10 s preceding the DEET pulse. Response and peak magnitudes were calculated using traces smoothed by 5 frames and identifying the maximum value within the DEET pulse. Five sets of experiments were conducted over 3 days for a total of 41 wild-type worms and 49 worms expressing *str-217*.

In Fig. 4n, traces were bleach-corrected using a custom MATLAB script and then the 5% of frames with the lowest values were averaged to create F_0 . $\Delta F/F_0$ (%) was calculated³⁸ by $(F - F_0)/F_0$. The heat-map traces in Fig. 4n were also smoothed by 5 frames, such that each data point (n) is the running average of $n - 2$, $n - 1$, n , $n + 1$ and $n + 2$. Peak $\Delta F/F_0$ (%) in Fig. 4o reflects the maximum value of $\Delta F/F_0$ (%) during the DEET pulse.

Chemotaxis tracking and analysis. Between 8 and 20 adult hermaphrodites were first transferred to an empty NGM plate and then 4–15 were transferred to an assay plate to minimize bacterial transfer. Worms were then placed in the centre on either a 0.15% DEET–agar or solvent–agar plate, and their movement was recorded for 60 min at 3 frames/s with 6.6 MP PL-B781F CMOS camera (PixeLINK) and Streampix software. Assays were carried out at room temperature, between

12:00 and 20:00 Eastern Time, and lit from below. Worm trajectories were extracted by a custom MATLAB (MathWorks) script¹⁵, and discontinuous tracks were then manually linked. Tracks were discarded if the worm moved less than two body lengths from its origin over the course of the 60-min trial. If a worm came within 1 cm of the isoamyl alcohol stimulus, the track was truncated to remove information from worms immobilized at the odorant source because of the addition of sodium azide.

ADL optogenetic stimulation. L4 worms expressing an *Psrh-220::ReaChR27* array or array-negative worms from the same plate were raised overnight in the dark on an NGM plate freshly seeded with 100 μ l of 10× concentrated *E. coli* (OP50 strain) with or without 50 μ M all-*trans* retinal (CID: 720648, Sigma-Aldrich, R2500), which is required for ReaChR-induced activity. The next day, adult hermaphrodites were first transferred to an empty NGM plate and then 4–15 worms were transferred to the 10-cm circular assay plate to minimize bacterial transfer. Videos were recorded for 26 min at 3 frames/s with a 1.3 MP PL-A741 camera (PixeLINK) and Streampix software. Blue light pulses were delivered with an LED (455 nm, 45 μ W/mm², Mightex) controlled with a custom MATLAB script^{15,39}. Worms were exposed to normal light for 120 s, before exposure to 6 repetitions of blue light (10 Hz strobing) for 120 s, and 120 s of recovery (LED off). Worm trajectories were extracted by a custom MATLAB script³⁹. Pausing events were extracted, and all pauses ≥ 3 frames (1 s) were used for further analysis. Pauses were classified as ‘on’ if any frame included light illumination. A pause that began just before illumination began—but remained paused while the illumination occurred—was considered an ‘on’ pause; any pauses that began during light illumination were also considered to be ‘on’. All other pauses were classified as ‘off’ pauses. In the analysis in Fig. 5j, we took an average pause length for all on and off pauses for each worm and pooled all of the worms on each plate. To control for any baseline differences between worms and experiment-to-experiment variation, we examined the increase in average pause length in Fig. 5k.

Phylogenetic analysis of *str-217* in wild isolates. Data were obtained from CeNDR¹⁹. Only predicted deletions in exons or missense changes of high confidence were included (Supplementary Data).

Statistical analysis. R v.3.3.2 was used for all statistical analysis. Additionally, qqPlots were evaluated before performing ANOVAs. For the analysis of optogenetic experiments, a Levene’s test identified heteroscedasticity in these data that was addressed with a boxcox translation. Imaging data from AWC neurons were similarly boxcox-translated and transformed to adjust for the rightward skew.

Strains. Detailed genotypes of all *C. elegans* strains and their sources are provided in Supplementary Data.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All scripts and graphed data with the exception of raw video files are available in the Supplementary Data. Raw video files are available on request from the corresponding author.

- Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
- Stiernagle, T. in *WormBook* (ed. The *C. elegans* Research Community, WormBook) <https://www.doi.org/10.1895/wormbook.1.101.1> (2006).
- Mello, C. & Fire, A. DNA transformation. *Methods Cell Biol.* **48**, 451–482 (1995).
- Arribere, J. A. et al. Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in *Caenorhabditis elegans*. *Genetics* **198**, 837–846 (2014).
- Hart, A. C. (ed.) in *WormBook* (ed. The *C. elegans* Research Community, WormBook) <https://www.doi.org/10.1895/wormbook.1.87.1> (2006).
- Sarin, S. et al. Analysis of multiple ethyl methanesulfonate-mutagenized *Caenorhabditis elegans* strains by whole-genome sequencing. *Genetics* **185**, 417–430 (2010).
- Zuryn, S., Le Gras, S., Jamet, K. & Jarriault, S. A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186**, 427–430 (2010).
- Kutscher, L. M. & Shaham, S. in *WormBook* (ed. The *C. elegans* Research Community, WormBook) <https://doi.org/10.1895/wormbook.1.167.1> (2014).
- Larsch, J. et al. A circuit for gradient climbing in *C. elegans* chemotaxis. *Cell Rep.* **12**, 1748–1760 (2015).
- Jang, H. et al. Dissection of neuronal gap junction circuits that regulate social behavior in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **114**, E1263–E1272 (2017).
- Larsch, J., Ventimiglia, D., Bargmann, C. I. & Albrecht, D. R. High-throughput imaging of neuronal activity in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **110**, E4266–E4273 (2013).
- Gordus, A., Pokala, N., Levy, S., Flavell, S. W. & Bargmann, C. I. Feedback from network states generates variability in a probabilistic olfactory circuit. *Cell* **161**, 215–227 (2015).

Coherent encoding of subjective spatial position in visual cortex and hippocampus

Aman B. Saleem^{1,2,5*}, E. Mika Diamanti^{1,3,5}, Julien Fournier¹, Kenneth D. Harris^{4,6} & Matteo Carandini^{1,6}

A major role of vision is to guide navigation, and navigation is strongly driven by vision^{1–4}. Indeed, the brain's visual and navigational systems are known to interact^{5,6}, and signals related to position in the environment have been suggested to appear as early as in the visual cortex^{6,7}. Here, to establish the nature of these signals, we recorded in the primary visual cortex (V1) and hippocampal area CA1 while mice traversed a corridor in virtual reality. The corridor contained identical visual landmarks in two positions, so that a purely visual neuron would respond similarly at those positions. Most V1 neurons, however, responded solely or more strongly to the landmarks in one position rather than the other. This modulation of visual responses by spatial location was not explained by factors such as running speed. To assess whether the modulation is related to navigational signals and to the animal's subjective estimate of position, we trained the mice to lick for a water

reward upon reaching a reward zone in the corridor. Neuronal populations in both CA1 and V1 encoded the animal's position along the corridor, and the errors in their representations were correlated. Moreover, both representations reflected the animal's subjective estimate of position, inferred from the animal's licks, better than its actual position. When animals licked in a given location—whether correctly or incorrectly—neural populations in both V1 and CA1 placed the animal in the reward zone. We conclude that visual responses in V1 are controlled by navigational signals, which are coherent with those encoded in hippocampus and reflect the animal's subjective position. The presence of such navigational signals as early as a primary sensory area suggests that they permeate sensory processing in the cortex.

To characterize the influence of spatial position on the responses of area V1, we took mice expressing the calcium indicator GCaMP6 in

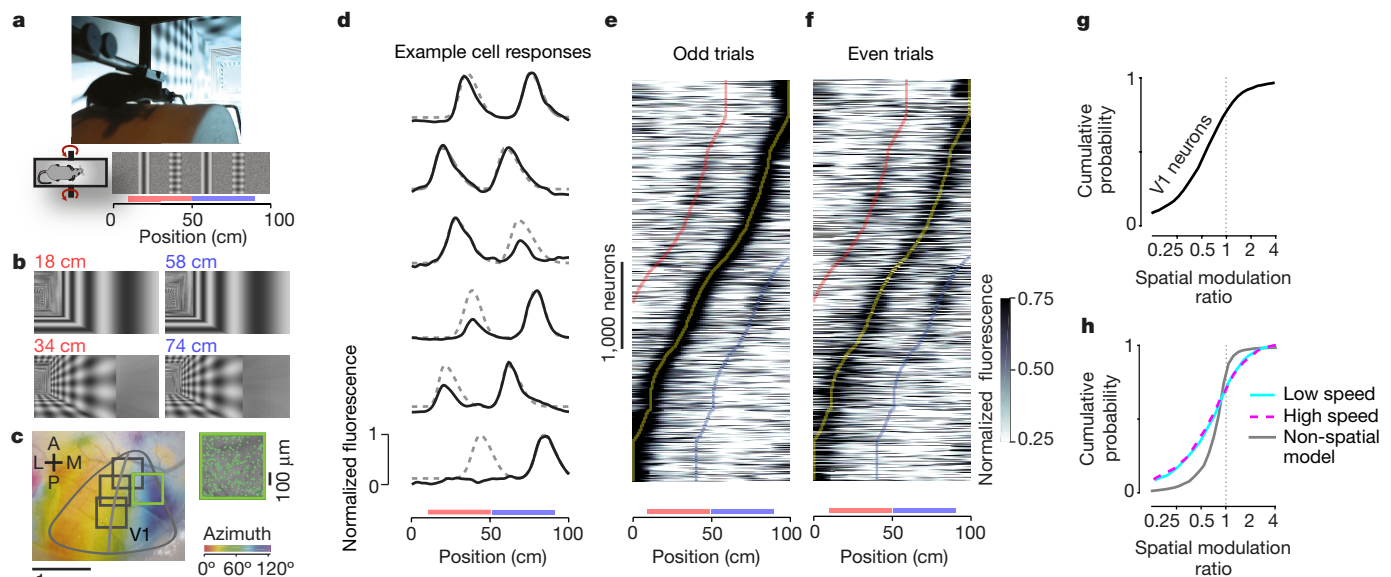


Fig. 1 | Responses in V1 are modulated by spatial position. **a**, Mice ran on a cylindrical treadmill to navigate a virtual corridor. The corridor had two landmarks that repeated after 40 cm, creating visually matching segments (red and blue bars). **b**, Screenshots showing the right half of the corridor at pairs of positions 40 cm apart. **c**, Example retinotopic map of the cortical surface. Grey curve shows the border of V1. Squares denote the field of view in two-photon imaging sessions targeted to medial V1 (inset shows the field with green frame). We analysed responses from neurons with receptive field centres greater than 40° azimuth (curve). **d**, Normalized response as a function of position in the corridor for six example V1 neurons. Dotted lines show predictions, assuming identical responses in matching segments of the corridor. **e**, Normalized response as a function of position, obtained from odd trials, for 4,958 V1 neurons.

Neurons are ordered by the position of their maximum response. **f**, As in **e** for even trials. Curves indicate preferred position (yellow) and preferred position ± 40 cm (blue and red). **g**, Cumulative distribution of the spatial modulation ratio in even trials: response at non-preferred position (40 cm from peak response) divided by response at preferred position for cells with responses within the visually matching segments (median \pm m.a.d., 0.61 ± 0.31 ; significantly less than 1, $P < 10^{-104}$, $n = 2,422$, Wilcoxon two-sided signed rank test). **h**, As in **g**, stratifying the data by running speed and considering a model without spatial selectivity, the non-spatial model. The curves corresponding to low (cyan) and high (purple) speeds overlap and appear as a single dashed curve ($P = 0.21$, Wilcoxon two-sided signed rank test). Grey curve, spatial modulation ratios from a non-spatial model considering visual and behavioural factors (Extended Data Fig. 7).

¹UCL Institute of Ophthalmology, University College London, London, UK. ²Department of Experimental Psychology, University College London, London, UK. ³CoMPLEX, Department of Computer Science, University College London, London, UK. ⁴UCL Institute of Neurology, University College London, London, UK. ⁵These authors contributed equally: Aman B. Saleem, E. Mika Diamanti. ⁶These authors jointly supervised this work: Kenneth D. Harris, Matteo Carandini. *e-mail: aman.saleem@ucl.ac.uk

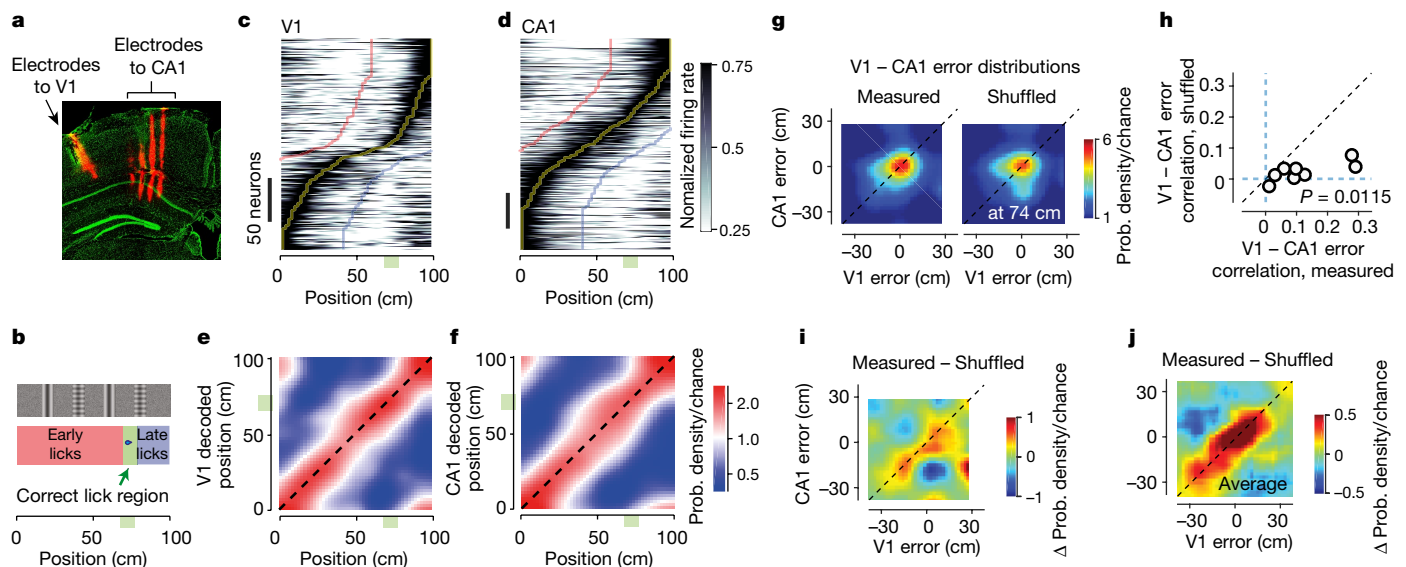


Fig. 2 | V1 and CA1 neural populations represent spatial positions in the virtual corridor and make correlated errors. **a**, Example of reconstructed electrode tracks (red: DiI); green shows cells labelled with DAPI. Panel shows tracks from one array (four shanks) in CA1, and a second electrode (one shank) in V1. **b**, In the task, water was delivered when mice licked in a reward zone (green area). **c**, Normalized activity as a function of position in the corridor, for 226 V1 neurons (8 sessions). Neurons are ordered by the position of their maximum response. Curves indicate preferred position (yellow) and preferred position ± 40 cm (blue and red). **d**, Similar plot for CA1 place cells (334 neurons; 8 sessions). **e**, Density map showing the distribution of position decoded from the activity of simultaneously recorded V1 neurons (y-axis) as a function of the animal's position (x-axis), averaged across recording sessions ($n = 8$),

and considering only correct trials. The red diagonal stripe indicates accurate estimation of position. **f**, Similar plot for CA1 neurons. **g**, Density map showing the joint distribution of position decoding errors from V1 and CA1 in one example session at one position (74 cm; left), together with a similar analysis on data shuffled while preserving the correlation due to running speed and position (right). **h**, Pearson's correlation coefficient of decoding errors in V1 and CA1 for each recording session ($n = 3,800$; 21,000 time points), against similar analysis of shuffled data. Correlations are above shuffling control ($P = 0.0115$, two-sided t -test, $n = 8$ sessions). **i**, Difference between joint distribution of V1 and CA1 decoded position and shuffled control, for the example in **g**. **j**, Difference between joint density map of V1 and CA1 decoded position, and shuffled control, averaged across positions ($n = 50$) and sessions ($n = 8$).

excitatory cells and placed them in a corridor in virtual reality (Fig. 1a). The corridor had a pair of landmarks (a grating and a plaid) that repeated twice, thus creating two visually matching segments 40 cm apart (Fig. 1a, b; Extended Data Fig. 1). We identified V1 using the retinotopic map measured using wide-field imaging (Fig. 1c). We then used a two-photon microscope to view medial V1, focusing our analysis on neurons with receptive field centres more lateral than 40° azimuth (Fig. 1c), which were driven as the mouse passed the landmarks. As expected, given the repetition of visual scenes in the two segments of the corridor, some V1 neurons had a response profile with two equal peaks 40 cm apart (Fig. 1d). Other V1 neurons, however, responded differently to the same visual stimuli in the two segments (Fig. 1d). These results indicate that visual activity in V1 can be strongly modulated by an animal's position in an environment.

This modulation of visual responses by spatial position occurred in the majority of V1 neurons (Fig. 1e–g). We imaged 8,610 V1 neurons across 18 sessions in 4 mice and selected 4,958 neurons with receptive field centres beyond 40° azimuth and reliable firing along the corridor (see Methods). We divided the trials in half, and used the odd-numbered trials to find the position at which each neuron fired maximally. The resulting representation reveals a striking preference of V1 neurons for spatial position (Fig. 1e), with most neurons giving stronger responses in one position (preferred position) than in the visually matching position 40 cm away (non-preferred position). To avoid circularity, we quantified this preference on the other half of the data (the even-numbered trials) and found that the preference for position was robust (Fig. 1f). Indeed, among the neurons that responded when the mouse traversed the visually matching segments ($n = 2,422$), the responses at the non-preferred position were markedly smaller than at the preferred position (Fig. 1g; Extended Data Fig. 2). We defined a spatial modulation ratio for each cell as the ratio of responses at the two visually matching positions (non-preferred/preferred, in the even trials). The median spatial modulation ratio was 0.61 ± 0.31 (\pm median

absolute deviation, m.a.d.), significantly less than 1 ($P < 10^{-104}$, Wilcoxon two-sided signed rank test). Neurons preferred the first or second sections in similar proportions (49% versus 51%), making it unlikely that a global factor such as visual adaptation could explain their preference.

The modulation of V1 responses by spatial position could not be explained by visual factors. To confirm that the receptive fields of most neurons saw similar stimuli in the two visually matching locations, we ran a model of receptive field responses (a simulation of V1 complex cells) on the sequences of images. As expected, this model generated spatial modulation ratios close to 1 (0.97 ± 0.17 , Extended Data Fig. 3). We next asked whether the different responses seen in the two locations could be due to differences in images far outside the receptive field, particularly the end (grey) wall of the corridor. To test this, we placed two additional mice in a modified virtual reality environment, in which the two sections of the corridor were pixel-to-pixel identical (Extended Data Fig. 4). The spatial modulation ratio was again overwhelmingly less than 1 (0.62 ± 0.26 ; $P < 10^{-81}$; $n = 1,044$ neurons), confirming that spatial modulation of V1 responses could not be explained by distant visual cues.

Spatial modulation of V1 responses could also not be explained by running speed, deviations in pupil position and diameter, or reward. Given that V1 neurons are influenced by running speed and visual speed^{8,9}, their different responses in visually matching segments of the corridor could reflect speed differences. To control for this, we stratified the data according to three running speed ranges (low, medium, or high; Extended Data Fig. 5). Even within a group (medium speed), the spatial modulation ratio was substantially below 1 (0.47 ± 0.22 ; $P < 10^{-33}$). Moreover, the spatial ratio of responses was identical at low and high speeds (Fig. 1h). We could also exclude a role of reward or deviations in pupil position and size, as the spatial modulation ratio was markedly below 1 even in sessions during which the animals ran without a reward (0.57 ± 0.37 ; $P < 10^{-14}$), or when there were no changes in pupil

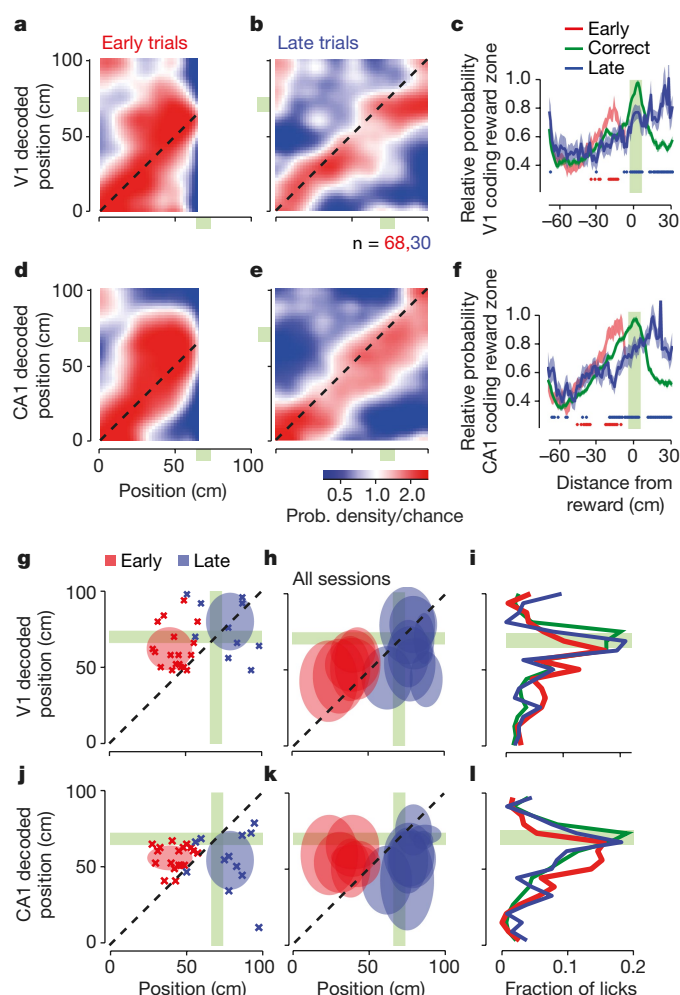


Fig. 3 | Positions encoded by visual cortex and hippocampus correlate with animal's spatial decisions. **a**, Distribution of positions decoded from the V1 population, as a function of the animal's actual position, on trials in which mice licked early. The decoder was trained on separate trials during which mice licked in the correct position. **b**, Same plot for trials during which mice licked late. **c**, The average decoded probability that the mouse is in the reward zone, as a function of distance from the reward. The curve for early trials (red) peaks before the reward zone, whereas the curve for late trials (blue) peaks after it, consistent with V1 activity reflecting subjective position rather than actual position. Probabilities were normalized relative to the probability of being in the reward zone in the correct trials (green). Red dots, positions at which the decoded probability of being in the reward zone differed significantly between early and correct trials ($P < 0.05$, two-sample two-sided t -test). Blue dots: same, for correct versus late trials. Shaded regions indicate mean \pm s.e.m., $n = 68$ early trials (red), 334 correct trials (green), and 30 late trials (blue). **d–f**, Same as **a–c**, for decoding using the population of CA1 neurons. **g**, Position decoded from V1 activity as a function of mouse position, in an example session. Crosses show positions when the animal licked during early (red) or late trials (blue). Late trials can include some early licks. These distributions (mean \pm s.d.) are summarized as shaded ovals for early trials (red, $n = 20$ licks) and late trials (blue, $n = 12$ licks). Green regions mark the reward zone. **h**, Summary distributions for all sessions ($n = 8$). **i**, Fraction of licks as a function of distance from reward location in positions decoded from V1 activity. **j–l**, Same as in **g–i**, for CA1 neurons.

size (0.63 ± 0.33 , $P < 10^{-45}$) or pupil position (0.63 ± 0.33 , $P < 10^{-27}$; Extended Data Fig. 6). To assess the joint contribution of visual, task-related, and position variables, we developed three prediction models (Extended Data Fig. 7). The first depended only on the visual scenes, which repeat twice, and on trial onset and offset, which introduce transients (visual model). The second additionally depended on running speed, reward times, pupil size, and eye position (non-spatial

model). The third, in addition, allowed responses to differ in amplitude in the matching segments (spatial model). Only the last model could fit the activity of cells with unequal peaks, thus matching the spatial modulation ratios seen in the data (Extended Data Fig. 7c, d). By contrast, the first two models predicted spatial modulation ratios closer to 1 (Fig. 1h; Extended Data Fig. 7c, d).

Having established that V1 responses are modulated by spatial position, we next investigated whether the underlying modulatory signals reflect the spatial position encoded in the brain's navigational systems (Fig. 2). We recorded simultaneously from V1 and hippocampal area CA1 using two 32-channel electrodes (Fig. 2a). To gauge a mouse's estimate of position, we trained the mice to lick a spout for water reward upon reaching a specific region of the corridor (Fig. 2b; Supplementary Video 1; Extended Data Fig. 8). All four mice (wild type) learned to perform this task with more than 80% accuracy and relied strongly on vision: performance persisted when we changed the gain relating wheel rotation to progression in the corridor^{3,10} and performance decreased when we lowered visual contrast (Extended Data Fig. 8).

Many neurons in both visual cortex and hippocampus had place-specific response profiles, thus encoding the mouse's spatial position (Fig. 2c–f). Consistent with our observations from two-photon imaging, V1 neurons responded more strongly in one of the two visually matching segments of the corridor (Fig. 2c, Extended Data Fig. 9c). In turn, hippocampal CA1 neurons exhibited place fields^{3,10,11}, responding in a single corridor location (Fig. 2d, Extended Data Fig. 9a–c). Therefore, responses in both V1 and CA1 encoded the position of the mouse in the environment, with no ambiguity between the two visually matching segments. Indeed, an independent Bayes decoder was able to read out the mouse's position from the activity of neurons recorded from V1 (33 ± 17 neurons per session, $n = 8$ sessions; Fig. 2e) or from CA1 (42 ± 20 neurons per session, $n = 8$ sessions; Fig. 2f).

Furthermore, when the visual cortex and hippocampus made errors in estimating the mouse's position, these errors were correlated with each other (Fig. 2g, h). The distributions of errors in position decoded from V1 and CA1 peaked at zero (Fig. 2g) but were significantly correlated (Fig. 2h; $\rho = 0.125$, $P = 0.0129$, two-sided t -test, $n = 8$). In principle, this correlation could arise from a common modulation of both regions by behavioural factors such as running speed, which affects responses of both visual cortex^{8,9} and hippocampus^{12–14}. To isolate the effect of speed, we shuffled the data between time points while preserving the relationship between speed and position (see Supplementary Methods). After shuffling, the correlation between decoding errors in V1 and CA1 decreased substantially from 0.125 to 0.022 ($P = 0.0115$; Fig. 2g, h). Moreover, when we subtracted the shuffled distribution from the original joint distributions, the residual decoding errors were distributed along the diagonal (Fig. 2i, j), indicating that representations in V1 and CA1 are more correlated than expected from common speed modulation. This correlation could also not be explained by common encoding of behavioural factors such as licking (Extended Data Fig. 9d–f). Indeed, a prediction of V1-encoded position from all external variables (true position, running speed, licks and rewards) could still be improved by the position decoded from CA1 activity (Extended Data Fig. 10).

We next tested whether the spatial position encoded by V1 and CA1 relates to the mouse's subjective estimate of position (Fig. 3a–f). CA1 activity is influenced by the performance of navigation tasks^{15–18}, and may reflect the animal's subjective position more than its actual position^{15,17,19}. We assessed a mouse's subjective estimate of position from the location of its licks. We divided trials into three groups: early trials, in which too many licks (usually 4–6) occurred before the reward zone, causing the trial to be aborted; correct trials, during which one or more licks occurred in the reward zone; and late trials, in which the mouse missed the reward zone and licked afterwards. To understand how spatial representations in V1 and CA1 related to this behaviour, we trained the Bayesian decoder on the activity measured in correct trials, and analyzed the likelihood of decoding different positions in the three types of trial. Decoding performance in early and late trials

showed systematic deviations: in early trials, V1 and CA1 overestimated the animal's progress along the corridor (deviation above the diagonal, Fig. 3a, d), whereas in late trials they underestimated it (deviation below the diagonal, Fig. 3b, e). Accordingly, the probability of being in the reward zone, predicted from both CA1 and V1, peaked before the reward zone in early trials and after it in late trials (Fig. 3c, f). These consistent deviations suggest that the representations of position in V1 and CA1 correlate with the animal's decisions to lick and thus reflect its subjective estimate of position.

The licks provide an opportunity to gauge when the mouse's subjective estimate of position lies in the reward zone. If activity in V1 and CA1 reflects subjective position, it should place the animal in the reward zone whether the animal correctly licked in that zone or incorrectly licked earlier or later. To test this prediction, we decoded activity in V1 and CA1 at the time of licks. By definition, the distributions of licks in early, correct, and late trials were spatially distinct (Fig. 3g, h, j, k). However, when plotted as a function of decoded position, these distributions came into register over the reward zone, whether the decoding was done from V1 (Fig. 3g–i) or from CA1 (Fig. 3j–l). Thus, regardless of the animal's position, when a mouse licked for a reward, the activity of both V1 and CA1 indicated a position in the reward zone.

Together, these results indicate that visual responses in V1 are modulated by the same spatial signals as those represented in the hippocampus, and that these signals reflect the animal's subjective estimate of position. This modulation may become stronger as environments become familiar^{6,7}, perhaps contributing to the changes observed in V1 as animals learn behavioural tasks^{20–22}. The correlation between representations in V1 and CA1 may be due to feed-forward signals from vision or feedback signals from navigational systems. Although V1 and CA1 are not directly connected, they could share spatial signals through indirect connections^{23,24}; these could involve the retrosplenial, parietal, entorhinal, or prefrontal cortices, which are known to carry spatial information^{25,26}. Further insights into the nature of these signals could be obtained by modulating the relationship between actual position and distance run^{3,10} or time²⁷, and by investigating more natural 2D environments^{28–30}. In such environments, however, it would be difficult to control and repeat visual stimulation, which proved essential in our study. Our results show that signals related to an animal's own estimate of position appear as early as in primary sensory cortex. This observation suggests that the mouse cortex does not keep a firm distinction between navigational and sensory systems; rather, spatial signals may permeate cortical processing.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0516-1>.

Received: 15 December 2017; Accepted: 24 July 2018;

Published online 10 September 2018.

- Muller, R. U. & Kubie, J. L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* **7**, 1951–1968 (1987).
- Wiener, S. I., Korshunov, V. A., Garcia, R. & Berthoz, A. Inertial, subthalamic and landmark cue control of hippocampal CA1 place cell activity. *Eur. J. Neurosci.* **7**, 2206–2219 (1995).
- Chen, G., King, J. A., Burgess, N. & O'Keefe, J. How vision and movement combine in the hippocampal place code. *Proc. Natl Acad. Sci. USA* **110**, 378–383 (2013).
- Geiller, T., Fattahi, M., Choi, J.-S. & Royer, S. Place cells are more strongly tied to landmarks in deep than in superficial CA1. *Nat. Commun.* **8**, 14531 (2017).
- Ji, D. & Wilson, M. A. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100–107 (2007).
- Haggerty, D. C. & Ji, D. Activities of visual cortical and hippocampal neurons co-fluctuate in freely moving rats during spatial behavior. *eLife* **4**, e08902 (2015).
- Fiser, A. et al. Experience-dependent spatial expectations in mouse visual cortex. *Nat. Neurosci.* **19**, 1658–1664 (2016).

- Niell, C. M. & Stryker, M. P. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**, 472–479 (2010).
- Saleem, A. B., Ayaz, A., Jeffery, K. J., Harris, K. D. & Carandini, M. Integration of visual motion and locomotion in mouse visual cortex. *Nat. Neurosci.* **16**, 1864–1869 (2013).
- Ravassard, P. et al. Multisensory control of hippocampal spatiotemporal selectivity. *Science* **340**, 1342–1346 (2013).
- Harvey, C. D., Collman, F., Dombeck, D. A. & Tank, D. W. Intracellular dynamics of hippocampal place cells during virtual navigation. *Nature* **461**, 941–946 (2009).
- McNaughton, B. L., Barnes, C. A. & O'Keefe, J. The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Exp. Brain Res.* **52**, 41–49 (1983).
- Wiener, S. I., Paul, C. A. & Eichenbaum, H. Spatial and behavioral correlates of hippocampal neuronal activity. *J. Neurosci.* **9**, 2737–2763 (1989).
- Czurkó, A., Hirase, H., Csicsvari, J. & Buzsáki, G. Sustained activation of hippocampal pyramidal cells by 'space clamping' in a running wheel. *Eur. J. Neurosci.* **11**, 344–352 (1999).
- O'Keefe, J. & Speakman, A. Single unit activity in the rat hippocampus during a spatial memory task. *Exp. Brain Res.* **68**, 1–27 (1987).
- Lenck-Santini, P. P., Save, E. & Poucet, B. Evidence for a relationship between place-cell spatial firing and spatial memory performance. *Hippocampus* **11**, 377–390 (2001).
- Lenck-Santini, P.-P., Muller, R. U., Save, E. & Poucet, B. Relationships between place cell firing fields and navigational decisions by rats. *J. Neurosci.* **22**, 9035–9047 (2002).
- Hok, V. et al. Goal-related activity in hippocampal place cells. *J. Neurosci.* **27**, 472–482 (2007).
- Rosenzweig, E. S., Redish, A. D., McNaughton, B. L. & Barnes, C. A. Hippocampal map realignment and spatial learning. *Nat. Neurosci.* **6**, 609–615 (2003).
- Makino, H. & Komiyama, T. Learning enhances the relative impact of top-down processing in the visual cortex. *Nat. Neurosci.* **18**, 1116–1122 (2015).
- Poort, J. et al. Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron* **86**, 1478–1490 (2015).
- Jurjut, O., Georgieva, P., Busse, L. & Katzner, S. Learning enhances sensory processing in mouse V1 before improving behavior. *J. Neurosci.* **37**, 6460–6474 (2017).
- Witter, M. P. et al. Cortico-hippocampal communication by way of parallel parahippocampal-subicular pathways. *Hippocampus* **10**, 398–410 (2000).
- Wang, Q., Gao, E. & Burkhalter, A. Gateways of ventral and dorsal streams in mouse visual cortex. *J. Neurosci.* **31**, 1905–1918 (2011).
- Moser, E. I., Kropff, E. & Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* **31**, 69–89 (2008).
- Grieves, R. M. & Jeffery, K. J. The representation of space in the brain. *Behav. Processes* **135**, 113–131 (2017).
- Eichenbaum, H. Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* **15**, 732–744 (2014).
- Cushman, J. D. et al. Multisensory control of multimodal behavior: do the legs know what the tongue is doing? *PLoS One* **8**, e80465 (2013).
- Aronov, D. & Tank, D. W. Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system. *Neuron* **84**, 442–456 (2014).
- Chen, G., King, J. A., Lu, Y., Cacucci, F. & Burgess, N. Spatial cell firing during virtual navigation of open arenas by head-restrained mice. *eLife* **7**, e34789 (2018).

Acknowledgements We thank N. Burgess and B. Haider for helpful discussions, and C. Reddy, S. Schroeder, and M. Krumin for help with experiments. This work was funded by a Sir Henry Dale Fellowship, awarded by the Wellcome Trust/Royal Society (grant 200501) to A.B.S., EPSRC PhD award F500351/1351 to E.M.D., Human Frontier Science Program and EC Horizon 2020 grants to J.F. (grant 709030), the Wellcome Trust (grants 205093 and 108726) to M.C. and K.D.H., the Simons Collaboration on the Global Brain (grant 325512) to M.C. and K.D.H. M.C. holds the GlaxoSmithKline/Fight for Sight Chair in Visual Neuroscience.

Reviewer information Nature thanks M. Mehta and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed to the design of the study. A.B.S. carried out the electrophysiology experiments and E.M.D. the imaging experiments; A.B.S., E.M.D. and J.F. analysed the data. All authors wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0516-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0516-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.B.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

All experiments were conducted according to the UK Animals (Scientific Procedures) Act, 1986 under personal and project licenses issued by the Home Office following ethical review.

For simultaneous recordings in V1 and CA1, we used four C57BL/6 mice (all male, implanted at 4–8 weeks of age). For calcium imaging experiments, we used double or triple transgenic mice expressing GCaMP6 in excitatory neurons (5 females, 1 male, implanted at 4–10 weeks of age). The triple transgenic mice expressed GCaMP6 fast³¹ (Emx1-Cre;Camk2a-tTA;Ai93, 3 mice). The double transgenic mice expressed GCaMP6 slow³² (Camk2a-tTA;tetO-G6s, 3 mice). Because Ai93 mice may exhibit aberrant cortical activity³³, we used the GCaMP6 slow mice to validate the results obtained from the GCaMP6 fast mice. Additional tests³³ confirmed that none of these mice displayed the aberrant activity that is sometimes seen in Ai93 mice. No randomization or blinding was performed in this study. No statistical methods were used to predetermine sample size.

Virtual environment and task. The virtual reality environment was a corridor adorned with a white noise background and four landmarks: two grating stimuli oriented orthogonal to the corridor and two plaid stimuli (Fig. 1a). The corridor dimensions were $100 \times 8 \times 8$ cm, and the landmarks (8 cm wide) were centred 20, 40, 60 and 80 cm from the start of the corridor. The mice navigated the environment by walking on a custom-made polystyrene wheel (15 cm wide, 18 cm diameter). Movements of the wheel were captured by a rotary encoder (2,400 pulses per rotation, Kübler, Germany), and used to control the virtual reality environment presented on three monitors surrounding the animal, as previously described⁹. When the mouse reached the end of the corridor, it was placed back at the start of the corridor after a 3–5-s presentation of a grey screen. Trials longer than 120 s were timed out and were excluded from further analysis.

Mice used for simultaneous V1 and CA1 recordings ($n = 4$ animals, 8 sessions) were trained to lick in a specific region of the corridor, the reward zone. This zone was centred at 70 cm and was 8 cm wide. Trials in which the animals were not engaged in the task, that is, when they ran through the environment without licking, were excluded from further analysis. The animal was rewarded for correct licks with $\sim 2 \mu\text{l}$ water using a solenoid valve (161T010; Neptune Research, USA), and licks were monitored using a custom device that detected breaks in an infrared beam.

Mice used for calcium imaging ($n = 6$ animals, 25 recording sessions) ran the two versions of the virtual corridor, with no specific task.

In the standard version of the corridor, two of the mice (10 sessions) were motivated to run with water rewards: one mouse received rewards at random positions along the corridor and the other at the end of the corridor. To control for the effect of the reward on V1 responses, no reward was delivered to two other mice (8 sessions).

To ensure that the spatial modulation of V1 responses could not be explained by the end wall of the corridor being more visible in the second half than in the first half, two additional mice used for calcium imaging were trained in a modified version of the corridor, where visual scenes were strictly identical 40 cm apart (7 sessions). In this environment, mice ran the same distance as before (100 cm) and were also placed back at the start of the corridor after a 3–5-s presentation of a grey screen. The same four landmarks were also centred in the same positions as before. However, the corridor was extended to 200 cm length, repeating the same sequence of landmarks (Extended Data Fig. 4). The virtual reality software was modified to render only up to 70 cm ahead of the animal, ensuring the visual scenes were strictly identical in the sections between 10 and 50 cm and 50 and 90 cm; the white noise background also repeated with the same 40 cm periodicity. Prior to recording in the 200 cm corridor, mice were first exposed to 5 sessions in the 100 cm corridor, then placed in the 200 cm corridor and allowed to habituate to the new environment for another two or three sessions before the start of recordings.

Surgery and training. The surgical methods are similar to those described previously^{9,34}. In brief, a custom head-plate with a circular chamber (3–4 mm diameter for electrophysiology; 8 mm for imaging) was implanted on 4–10-week-old mice under isoflurane anaesthesia. For imaging, we performed a 4-mm craniotomy over the left visual cortex by repeatedly rotating a biopsy punch. The craniotomy was shielded with a double coverslip (4 mm inner diameter; 5 mm outer diameter). After 4 days of recovery, some mice were water restricted (>40 ml/kg/day) and were trained for 30–60 min, 5–7 days/week.

Mice used for simultaneous V1 and CA1 recordings were trained to lick selectively in the reward zone using a progressive training procedure. Initially, the animals were rewarded for running past the reward location on all trials. After this, we introduced trials in which the mouse was rewarded only when it licked in the rewarded region of the corridor. The width of the reward region was progressively narrowed from 30 cm to 8 cm across successive days of training. To prevent the animals from licking all across the corridor, trials were terminated early if the animal licked more than a certain number of times before the rewarded region. We reduced this number as the animals performed more accurately, typically reaching

a level of 4–6 licks by the time recordings were made. Once a sufficient level of performance was reached, we controlled on some (random) trials that the animal performed the task visually by measuring the performance when we decreased visual contrast or changed the distance to the reward zone (Extended Data Fig. 8). Training was carried out for 3–5 weeks. Animals were kept under light-shifted conditions (9 a.m. light off, 9 p.m. light on) and experiments were performed during the day.

Widefield calcium imaging. For widefield imaging we used a standard epillumination imaging system^{35,36} together with an sCMOS camera (pco.edge, PCO AG). A Leica $1.6\times$ Plan APO objective was placed above the imaging window and a custom black cone surrounding the objective was fixed on top of the headplate to prevent contamination from the monitors' light. The excitation light beam emitted by a high-power LED (465 nm LEX2-B, Brain Vision) was directed onto the imaging window by a dichroic mirror designed to reflect blue light. Emitted fluorescence passed through the same dichroic mirror and was then selectively transmitted by an emission filter (FF01-543/50-25, Semrock) before being focused by another objective (Leica 1.0 Plan APO objective) and finally detected by the camera. Images of 200×180 pixels, corresponding to an area of 6.0×5.4 mm, were acquired at 50 Hz.

To measure retinotopy we presented a 14° wide vertical window containing a vertical grating (spatial frequency 0.15 cycles per degree), and swept^{37,38} the horizontal position of the window over 135° of azimuth angle, at a frequency of 2 Hz. Stimuli lasted 4 s and were repeated 20 times (10 in each direction). We obtained maps for preferred azimuth by combining responses to the two stimuli moving in opposite directions, as previously described³⁷.

Two-photon imaging. Two-photon imaging was performed with a standard multiphoton imaging system (Bergamo II; Thorlabs) controlled by ScanImage⁴³⁹. A 970 nm laser beam, emitted by a Ti:sapphire laser (Chameleon Vision, Coherent), was targeted onto L2/3 neurons through a $16\times$ water-immersion objective (0.8 NA, Nikon). The fluorescence signal was transmitted by a dichroic beamsplitter and amplified by photomultiplier tubes (GaAsP, Hamamatsu). The emission light path between the focal plane and the objective was shielded with a custom-made plastic cone, to prevent contamination from the monitors' light. In each experiment, we imaged four planes set apart by 40 μm . Multiple-plane imaging was enabled by a piezo focusing device (P-725.4CA PIFOC, Physik Instrumente), and an electro-optical modulator (M350-80LA, Conoptics Inc.), which allowed us to adjust the laser power with depth. Images of 512×512 pixels, corresponding to a field of view of $500 \times 500 \mu\text{m}$, were acquired at a frame rate of 30 Hz (7.5 Hz per plane).

Pre-processing of raw imaging movies was done using the Suite2p pipeline⁴⁰ and involved: 1) image registration to correct for brain movement; 2) ROI extraction (that is, cell detection); and 3) correction for neuropil contamination. For neuropil correction, we used an established method^{41,42}. We used Suite2p to determine a mask surrounding each cell's soma, the 'neuropil mask'. The inner diameter of the mask was 3 μm and the outer diameter was $<45 \mu\text{m}$. For each cell we obtained a correction factor, α , by regressing the binned neuropil signal (20 bins in total) from the fifth percentile of the raw binned cell signal. For a given session, we obtained the average correction factor across cells. This average factor was used to obtain the corrected individual cell traces, from the raw cell traces and the neuropil signal, assuming a linear relationship. All correction factors fell between 0.7 and 0.9.

To manually curate the output of Suite2p, we used two criteria: one anatomical and one activity-dependent. One of the anatomical criteria in Suite2p is 'area', that is, mean distance of pixels from ROI centre, normalized to the same measure for a perfect disk. We used this criterion (area <1.04) to exclude ROIs that were likely to correspond to dendrites rather than somata. The activity-related criterion is the standard deviation of the cell trace, normalized to the standard deviation of the neuropil trace. We used this criterion to exclude ROIs whose activity was too small relative to the corresponding neuropil signal (typically with $\text{std}(\text{neuropil corrected trace})/\text{std}(\text{neuropil signal}) < 2$). We finally excluded cells that fired extremely seldom (once or twice within a 20 min session).

Pupil tracking. We tracked the eye of the animal using an infrared camera (DMK 21BU04.H, Imaging Source) and a zoom lens (MVL7000, Navitar) at 25 Hz. Pupil position and size were calculated by fitting an ellipsoid to the pupil for each frame using a custom software. X and Y positions of the pupil were derived from the centre of mass of the fitted ellipsoid.

Electrophysiological recordings. On the day before the first recording session, we made two 1-mm craniotomies, one over CA1 (1.0 mm lateral, 2.0 mm anterior from lambda), and a second one over V1 (2.5 mm lateral, 0.5 mm anterior from lambda). We covered the chamber using KwikCast (World Precision Instruments) and the mice were allowed to recover overnight. The CA1 probe was lowered until all shanks were in the pyramidal layer, which was identified by the increase in theta power (5–8 Hz) of the local field potential and an increase in the number of detected units. The V1 probe was lowered to a depth of $\sim 800 \mu\text{m}$. We waited ~ 30 min for the tissue to settle before starting the recordings. In two mice, we dipped the probes in red-fluorescent DiI (Fig. 2a). In these mice, we had only

one recording session. The other two mice underwent two and four recording sessions, respectively.

Offline spike sorting was carried out using the KlustaSuite⁴³ package, with automated spike sorting using KlustaKwik⁴⁴, followed by manual refinement using KlustaViewa⁴⁵. Hippocampal interneurons were identified by their spike time autocorrelation and excluded from further analysis. Only time points with running speeds greater than 5 cm/s were included in further analyses.

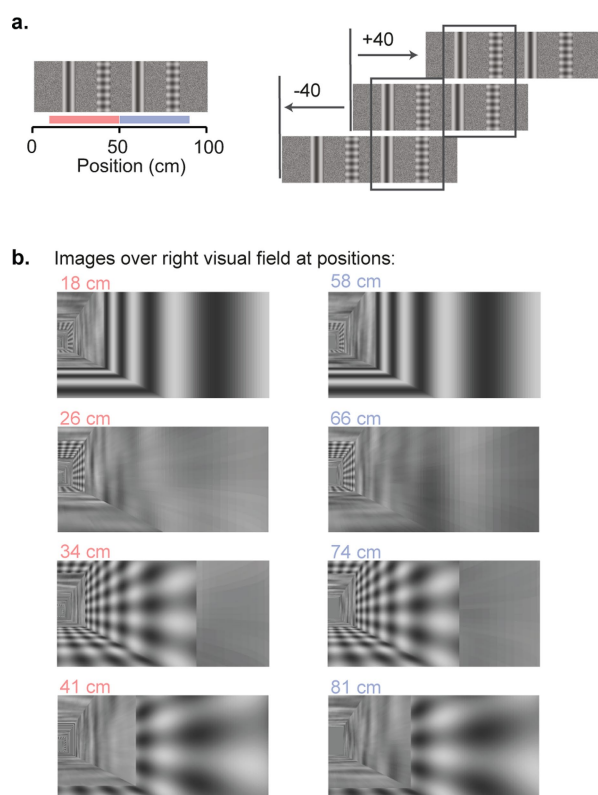
Data analysis and modelling methods. See Supplementary Methods for details of analysis and models.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

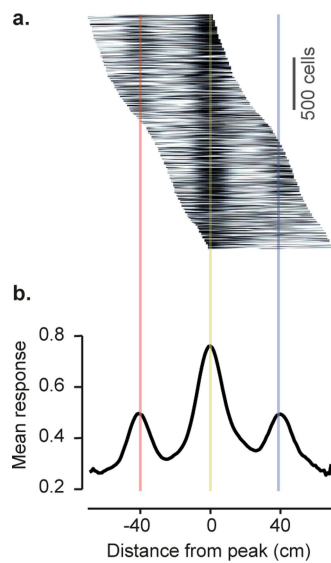
Code availability. The custom code from this study is available from the corresponding author upon reasonable request.

Data availability. The data from this study are available from the corresponding author upon reasonable request.

31. Madisen, L. et al. Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron* **85**, 942–958 (2015).
32. Wekselblatt, J. B., Flister, E. D., Piscopo, D. M. & Niell, C. M. Large-scale imaging of cortical dynamics during sensory perception and behavior. *J. Neurophysiol.* **115**, 2852–2866 (2016).
33. Steinmetz, N. A. et al. Aberrant cortical activity in multiple GCaMP6-expressing transgenic mouse lines. *eNeuro* <https://doi.org/10.1523/ENEURO.0207-17.2017> (2017).
34. Ayaz, A., Saleem, A. B., Schölvinck, M. L. & Carandini, M. Locomotion controls spatial integration in mouse visual cortex. *Curr. Biol.* **23**, 890–894 (2013).
35. Ratzlaff, E. H. & Grinvald, A. A tandem-lens epifluorescence microscope: hundred-fold brightness advantage for wide-field imaging. *J. Neurosci. Methods* **36**, 127–137 (1991).
36. Carandini, M. et al. Imaging the awake visual cortex with a genetically encoded voltage indicator. *J. Neurosci.* **35**, 53–63 (2015).
37. Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. *Neuron* **38**, 529–545 (2003).
38. Yang, Z., Heeger, D. J. & Seidemann, E. Rapid and precise retinotopic mapping of the visual cortex obtained by voltage-sensitive dye imaging in the behaving monkey. *J. Neurophysiol.* **98**, 1002–1014 (2007).
39. Polgruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
40. Pachitariu, M. et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. Preprint at <https://www.biorxiv.org/content/early/2017/07/20/061507> (2016).
41. Peron, S. P., Freeman, J., Iyer, V., Guo, C. & Svoboda, K. A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron* **86**, 783–799 (2015).
42. Dipoppa, M. et al. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron* **98**, 602–615.e8 (2018).
43. Rossant, C. et al. Spike sorting for large, dense electrode arrays. *Nat. Neurosci.* **19**, 634–641 (2016).
44. Kadir, S. N., Goodman, D. F. M. & Harris, K. D. High-dimensional cluster analysis with the masked EM algorithm. *Neural Comput.* **26**, 2379–2394 (2014).

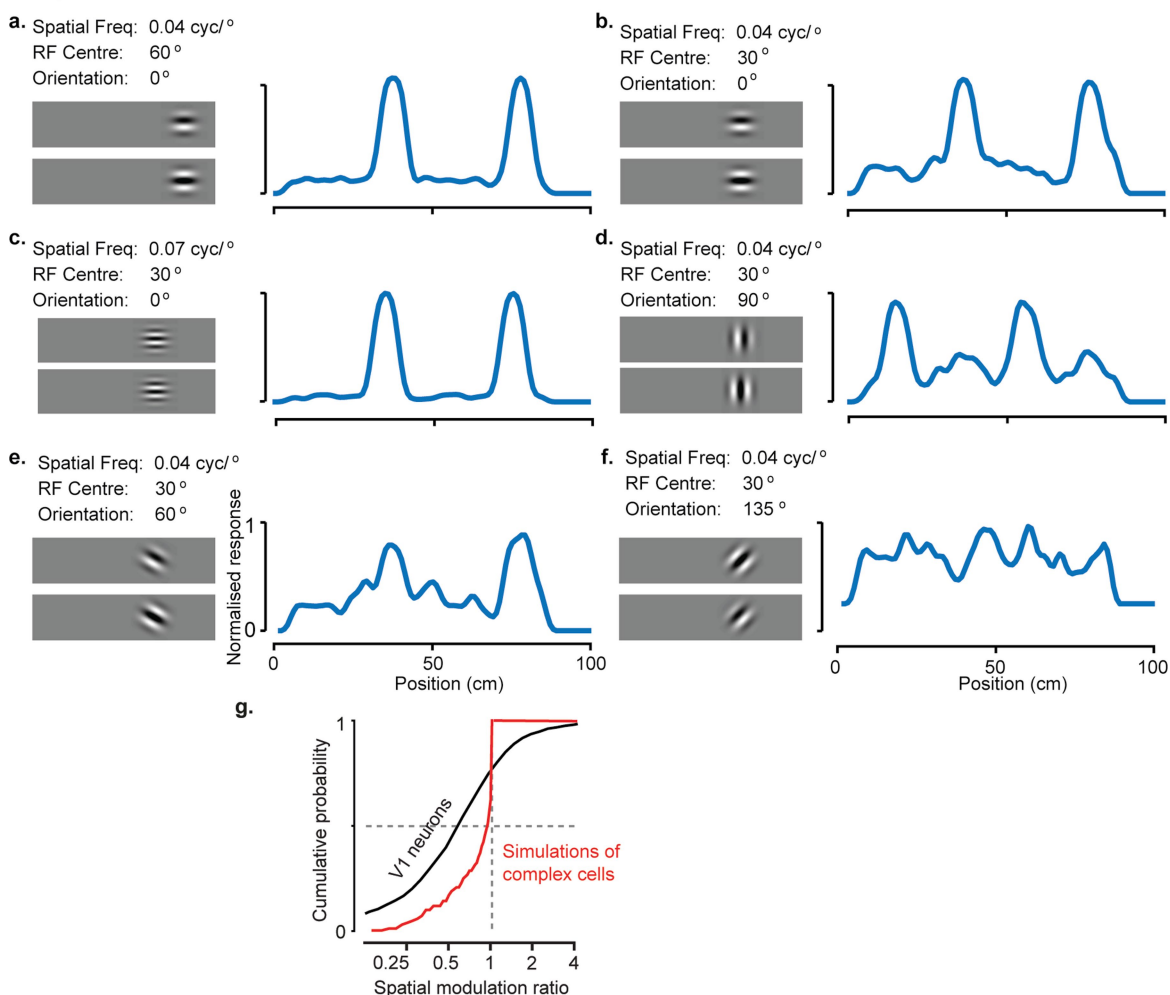


Extended Data Fig. 1 | Design of virtual environment with two visually matching segments. a, The virtual corridor had four prominent landmarks. Two visual patterns (grating and plaid) were repeated at two positions, 40 cm apart, to create two visually matching segments in the room, from 10 cm to 50 cm and from 50 cm to 90 cm (red and blue bars in the left panel), as illustrated in the right panel. **b,** Example screenshots of the right visual field displayed in the environment when the animal is at different positions. Each row displays screen images at positions approximately 40 cm apart.



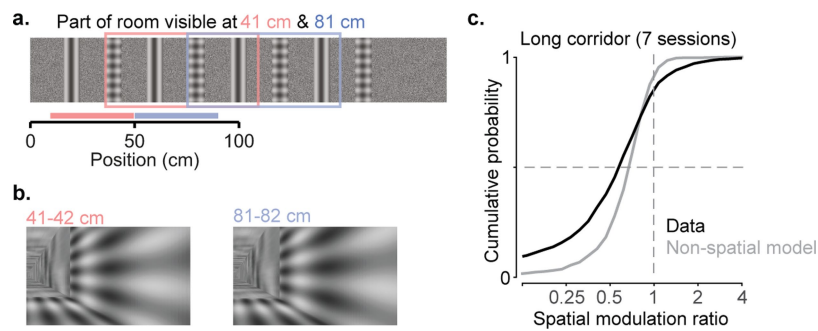
Extended Data Fig. 2 | Spatial averaging of visual cortical activity confirms the difference in response between visually matching locations. **a**, Mean response of V1 neurons as a function of the distance from the peak response location (2,422 cells with peak response between 15 and 85 cm along the corridor). To ensure that the average captured reliable, spatially specific responses, the peak response location for each cell was estimated only from odd trials, whereas the mean response was computed only from even trials. **b**, Population average of responses shown in **a**. Lower values of the side peaks compared to central peak indicate strong preference of V1 neurons for one segment of the corridor over the other visually matching segment (40 cm from peak response).

Complex cell simulations



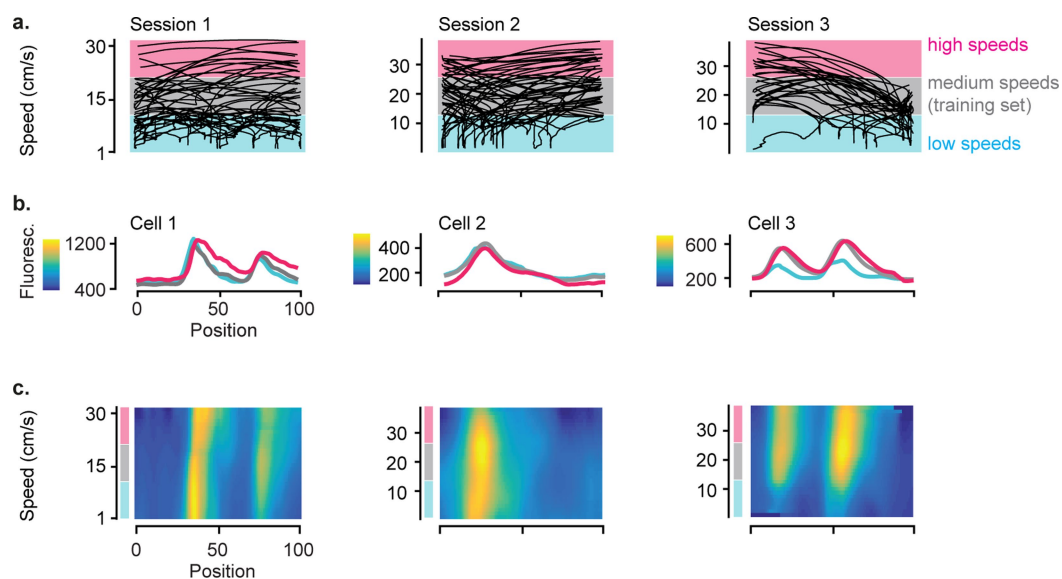
Extended Data Fig. 3 | Simulation of purely visual responses to position in VR. **a–f,** Responses of six simulated neurons with purely visual responses, produced by a complex cell model with varying spatial frequency, orientation, or receptive field location. The images on the left of each panel show the quadrature pair of complex cell filters; traces on the right show the cell's simulated response as a function of position in the virtual environment. Simulation parameters matched those that are commonly observed in mouse V1 (spatial frequency: 0.04, 0.05, 0.06 or 0.07 cycles per degree; orientation: uniform between 0° and 179° but with

twice as many cells for cardinal orientations; receptive field positions 40°, 50°, 70° and 80°, similar to the V1 neurons we considered for analysis. In rare cases (as in **f**) when the receptive fields do not match the features of the environment, there is little selectivity along the corridor. These cases lead to lower spatial modulation ratios. **g.** The spatial modulation ratios calculated for the complex cell simulations are close to 1 (0.97 ± 0.17), and different from the ratios calculated for V1 neurons. Black curve is the same as in Fig 1g.



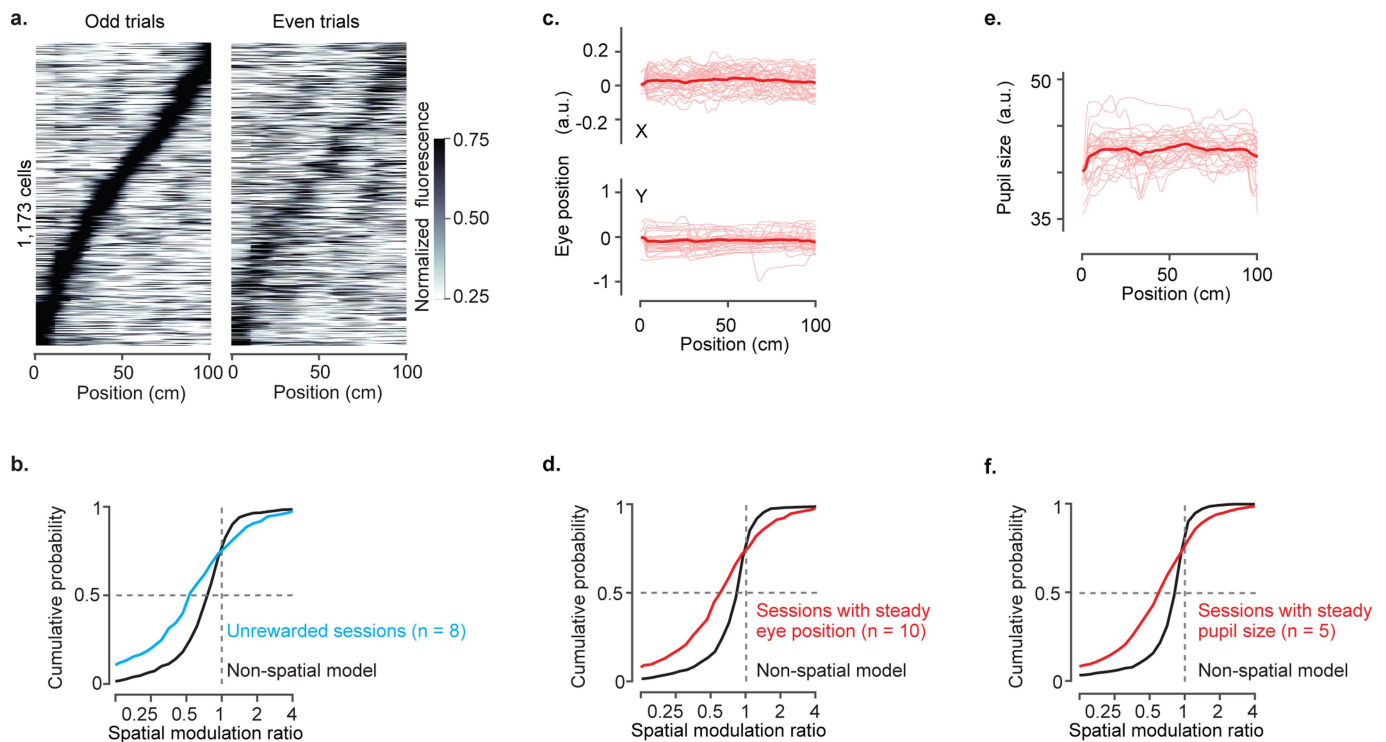
Extended Data Fig. 4 | The spatial modulation of V1 responses is not due to end-of-corridor visual cues. **a**, Diagram of the 200-cm virtual corridor, containing the same grating and plaid as the regular corridor, repeated four times instead of twice. **b**, Visual scenes from locations within the first 100 cm of the extended corridor, separated by 40 cm, are visually (pixel-to-pixel) identical. **c**, Cumulative distribution of the spatial modulation ratio across the two mice that were placed in the long corridor

(7 sessions, 2 mice; median \pm m.a.d: 0.62 ± 0.26 ; 1,044 neurons, black line). Grey line shows the spatial modulation ratio predicted by the non-spatial model (which predicts activity from the visual scene, trial onset and offset, speed, reward, pupil size and displacement from the central position of the eye; see Extended Data Fig. 7, non-spatial model). The two distributions are significantly different (two-sided Wilcoxon rank sum test; $P < 10^{-14}$).



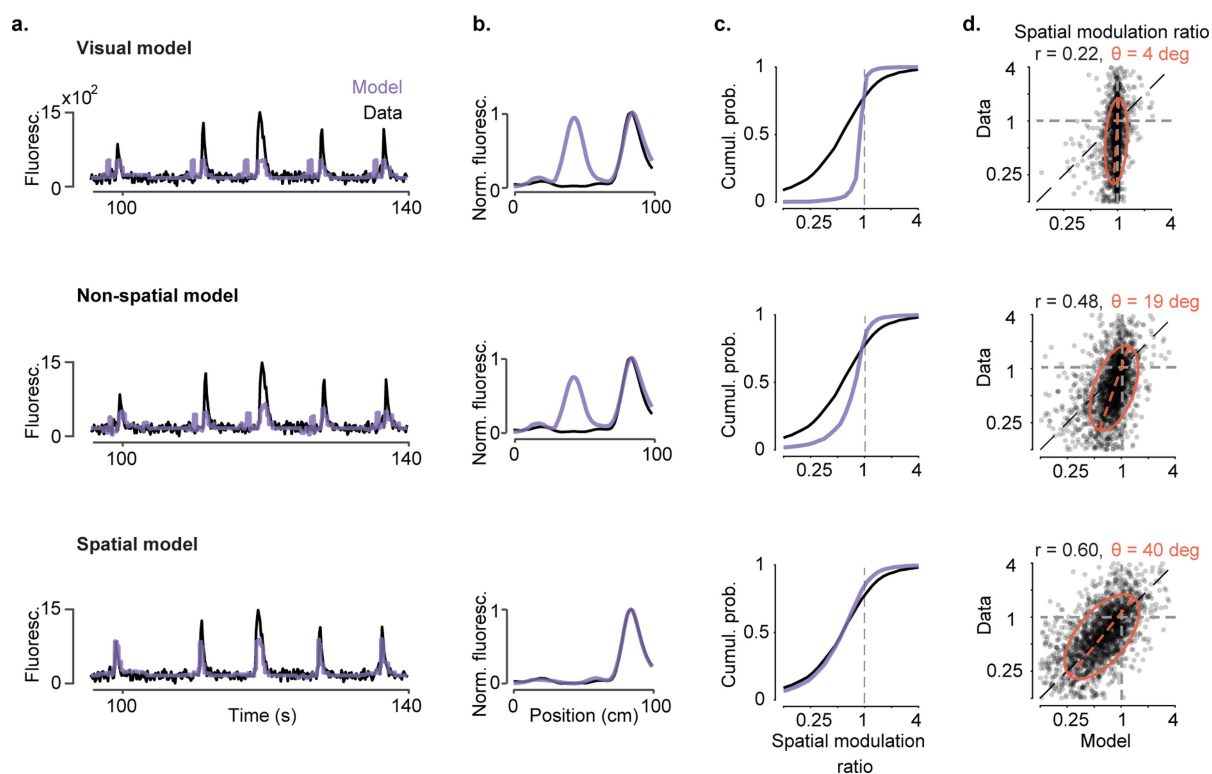
Extended Data Fig. 5 | The spatial modulation of V1 responses cannot be explained by speed. **a**, Speed–position plots for all single-trial trajectories in three example recording sessions. **b**, Response profile of example V1 cells in each session as a function of position in the corridor,

stratified for three speed ranges corresponding to the shading bands in **a**. **c**, Two-dimensional response profiles of the same example neurons showing activity as a function of position and running speed for speeds higher than 1 cm s^{-1} .



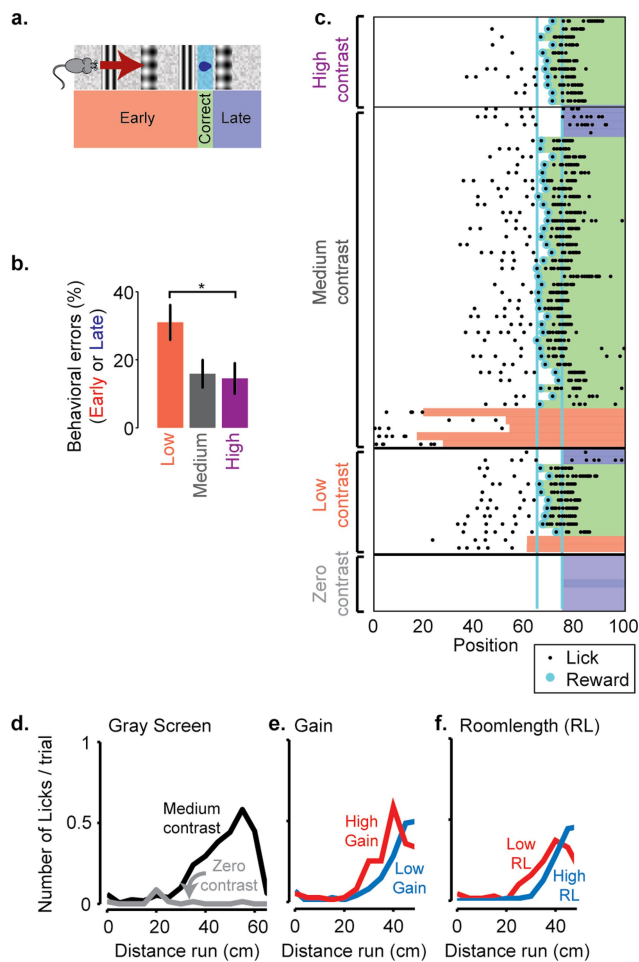
Extended Data Fig. 6 | The spatial modulation of V1 responses cannot be explained by reward, pupil position or diameter. **a**, Normalized response as a function of position in the virtual corridor, for sessions without reward (1,173 neurons). Data come from two out of four mice that ran the environment without reward (8 sessions, 2 mice). Responses in even trials (right) are ordered according to the position of maximum activity measured in odd trials (left). **b**, Distribution of spatial modulation ratio for unrewarded sessions (8 sessions; median \pm m.a.d. = 0.57 ± 0.37 ; cyan) and for modelled ratios obtained from the non-spatial model on the same sessions (black, see Extended Data Fig. 7). The two distributions are significantly different (two-sided Wilcoxon rank sum test; $P < 10^{-8}$). **c**, Pupil position as a function of location in the virtual corridor, for an example session with steady eye position. Sessions with steady eye positions were defined as those with no significant difference in eye

positions between visually matching positions 40 cm apart (with unpaired t -test, $P < 0.01$). Thin red curves: position trajectories on individual trials; thick curves, average. Top and bottom panels: x and y coordinates of the pupil, respectively. **d**, Distribution of spatial modulation ratio for sessions with steady eye position (10 sessions; median \pm m.a.d. = 0.63 ± 0.33 ; 1,154 neurons, red) and for modelled ratios obtained from the non-spatial model on the same sessions (black). The two distributions are significantly different (two-sided Wilcoxon rank sum test; $P < 10^{-14}$). **e**, Pupil size as a function of position for an example session with steady pupil size. **f**, Distribution of spatial modulation ratio for sessions with steady pupil size (5 sessions; median \pm m.a.d. = 0.63 ± 0.33 ; 1,069 neurons, red) and for modelled ratios obtained from the non-spatial model on the same sessions (black). The two distributions are significantly different (two-sided Wilcoxon rank sum test; $P < 10^{-13}$).



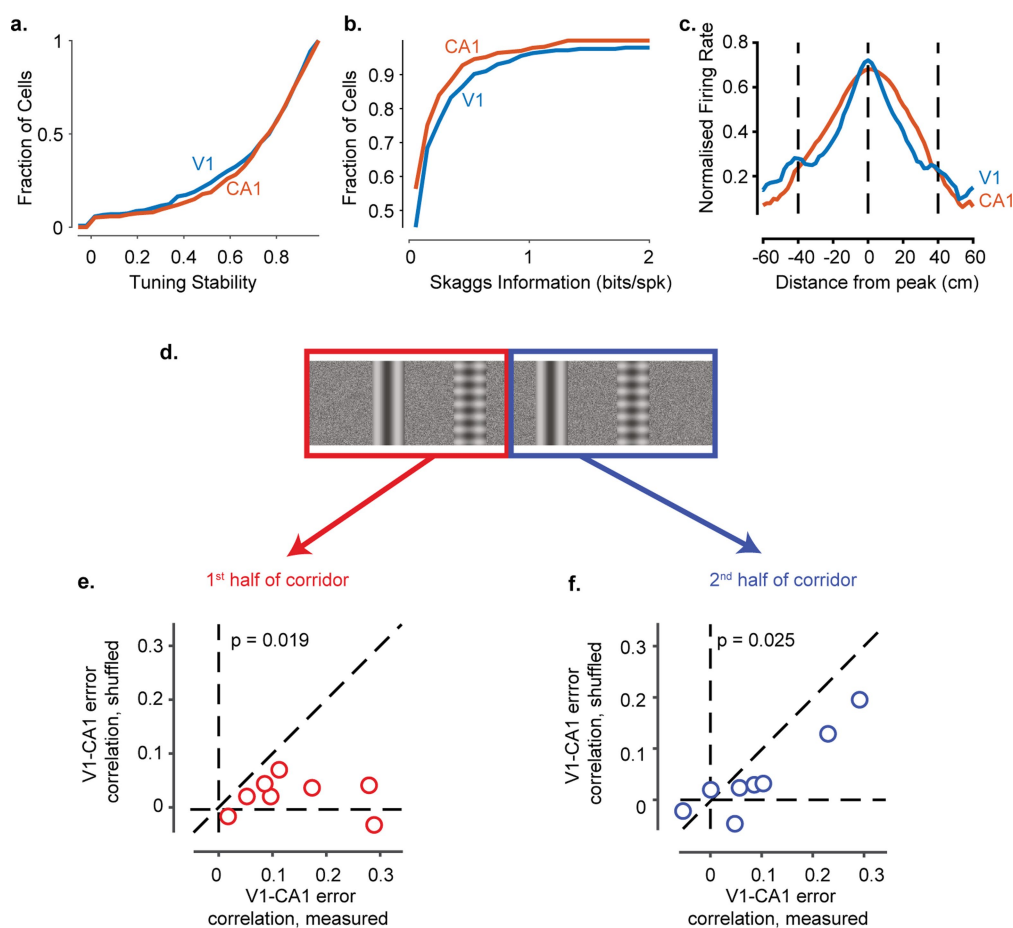
Extended Data Fig. 7 | Observed values of spatial modulation ratio can be modelled only using spatial position. **a, b,** We constructed three models to predict the activity of individual V1 neurons from successively larger sets of predictor variables. In the simplest, the visual model, activity is required to depend only on the visual scene visible from the mouse's current location, and is thus constrained to be a function of space that repeats in the visually matching section of the corridor. The second, non-spatial model, also includes modulation by behavioural factors that can differ within and across trials: speed, reward times, pupil size, and eye position. Because these variables can differ between the first and second halves of the track, modelled responses need no longer be exactly symmetrical; however, this model does not explicitly use space as a predictor. The final, spatial model, extends the previous model by also allowing responses to the two matching segments to vary in amplitude, thereby explicitly including space as a predictor. Example

single-trial predictions are shown as a function of time in **a**, together with measured fluorescence. Spatial profiles derived from these predictions are shown in **b**. **c**, Cumulative distributions of spatial modulation ratio for the three models (purple). For comparison, the black curve shows the ratio of peaks derived from the data (even trials) (median \pm m.a.d.: visual model, 0.99 ± 0.03 ; $P < 10^{-40}$, two-sided Wilcoxon rank sum test; non-spatial model, 0.83 ± 0.18 ; $P < 10^{-40}$; spatial model, 0.60 ± 0.27 ; $P = 0.09$, $n = 2,422$ neurons). **d**, Measured spatial modulation ratio versus predictions of the three models. Each point represents a cell; red ellipse represents best fit Gaussian, dotted line measures its slope. The purely visual model (top) does poorly, and is improved only slightly by adding predictions from speed, reward, pupil size, and eye position (middle). Adding an explicit prediction from space provides a much better match to the data (bottom). r , Pearson's correlation coefficient, $n = 2,422$ neurons; θ , orientation of the major axis of the fitted ellipsoid.



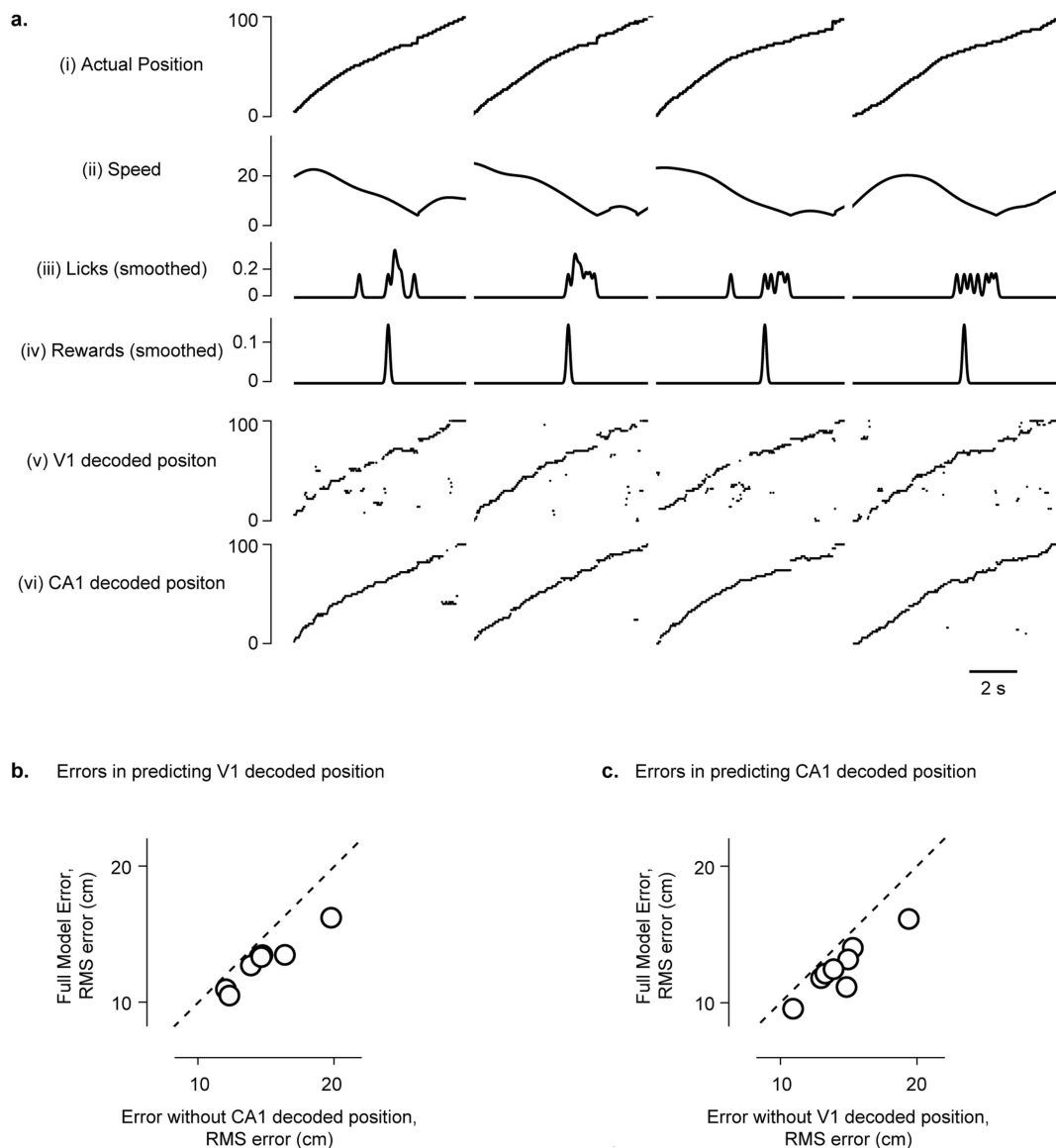
Extended Data Fig. 8 | Behavioural performance in the task.

a. Illustration of the virtual reality environment with four prominent landmarks, a reward zone, and the zones that define trial types: early, correct and late. **b.** Percentage of trials during which the animal makes behavioural errors, by licking either too early or too late at three different contrast levels: 18% (low), 60% (medium) or 72% (high). **c.** Illustration of performance on all trials of one example recording session. Each row represents a trial, black dots represent positions where the animal licked, and cyan dots indicate the delivery of a water reward. Coloured bars indicate the outcome of the trial (red, early; green, correct; blue, late). **d–f.** Successful performance relies on vision. **d.** The mouse did not lick when the room was presented at zero contrast. **e.** On some trials, we changed the gain between the animals' physical movement and movement in the virtual environment, thus changing the distance to the reward zone (high gain resulting in shorter distance), while visual cues remained in the same place. When plotted as a function of the distance run, the licks of the animal shifted, indicating that the animal was not relying simply on the distance travelled from the beginning of the corridor. **f.** If the position of the visual cues was shifted forward or back (high or low room length (RL)), the lick position shifted accordingly, indicating that the animals relied on vision to perform the task.



Extended Data Fig. 9 | Comparison of response properties between V1 and CA1 neurons and correlation of V1 and CA1 errors in the two halves of the environment. **a**, Cumulative distribution of the stability of V1 and CA1 response profiles. Tuning stability (the stability of responses) was computed as the correlation between the spatial responses measured from the first half and the second half of the trials. V1 and CA1 responses were highly stable within each recording session: the tuning stability was >0.7 for more than 60% of neurons in both V1 and CA1. **b**, Cumulative distribution of the Skaggs information (bits per spike) carried by V1 and CA1 neurons. Note that while V1 and CA1 neurons had comparable amounts of spatial information, this does not suggest that V1 represents space as strongly as CA1, because the Skaggs information metric mixes the influences of vision and spatial modulation. **c**, Normalized firing rate averaged across V1 or CA1 neurons as a function of distance from the peak response (similar to Extended Data Fig. 2b). Unlike CA1, the mean

response averaged from V1 neurons shows a second peak at ± 40 cm, consistent with the repetition of the visual scene. **d**, **e**, Pearson's correlation between position errors estimated from V1 and CA1 populations in the first half of the corridor (shown in **d**). Each point represents a behavioural session ($n = 8$ sessions); x-axis values represent measured correlations; y-axis values represent correlations calculated after having shuffled the data within the times where the speed was similar (similar to Fig. 2h). The occurrence of error correlations in the unshuffled data indicates that these correlations are not due to rewards (which did not occur in this half of the maze) or licks (which were rare, and the 100-ms periods surrounding the few that occurred were removed from analysis). The significance of the difference between the measured and shuffled correlations was calculated using a two-sided two-sample t -test. **f**, Similar to **e** for the second half of the corridor.



Extended Data Fig. 10 | Position decoded from CA1 activity helps to predict position decoded from V1 activity (and vice versa). **a.** To test whether the positions encoded in V1 and CA1 populations are correlated with each other beyond what would be expected from a common influence of other spatial and non-spatial factors, we used a random forests decoder (Tree Bagger implementation in MATLAB) to predict V1 or CA1 decoded positions from different predictors. We then tested whether the model prediction was further improved when we added the position decoded

from the other area as an additional predictor (that is, using the positions decoded from CA1 to predict V1 decoded positions and vice versa). **b.** Adding CA1 decoded position as an additional predictor improved the prediction of V1 decoded positions in every recording session (that is, reduced the prediction errors). V1 and CA1 decoded positions are thus correlated with each other beyond what can be expected from a common contribution of position, speed, licks and reward to V1 and CA1 responses. **c.** Same as **b** for predicting CA1 decoded position.

Mechanosensing by $\beta 1$ integrin induces angiocrine signals for liver growth and survival

Linda Lorenz^{1,2,3,11}, Jennifer Axnick^{1,11}, Tobias Buschmann^{1,4}, Carina Henning¹, Sofia Urner¹, Shentong Fang⁵, Harri Nurmi⁵, Nicole Eichhorst⁴, Richard Holtmeier^{6,7}, Kálmán Bódis^{3,8,9}, Jong-Hee Hwang^{8,9}, Karsten Müssig^{3,8,9}, Daniel Eberhard¹, Jörg Stypmann^{6,7}, Oliver Kuss^{3,10}, Michael Roden^{3,8,9}, Kari Alitalo⁵, Dieter Häussinger⁴ & Eckhard Lammert^{1,2,3,*}

Angiocrine signals derived from endothelial cells are an important component of intercellular communication and have a key role in organ growth, regeneration and disease^{1–4}. These signals have been identified and studied in multiple organs, including the liver, pancreas, lung, heart, bone, bone marrow, central nervous system, retina and some cancers^{1–4}. Here we use the developing liver as a model organ to study angiocrine signals^{5,6}, and show that the growth rate of the liver correlates both spatially and temporally with blood perfusion to this organ. By manipulating blood flow through the liver vasculature, we demonstrate that vessel perfusion activates $\beta 1$ integrin and vascular endothelial growth factor receptor 3 (VEGFR3). Notably, both $\beta 1$ integrin and VEGFR3 are strictly required for normal production of hepatocyte growth factor, survival of hepatocytes and liver growth. Ex vivo perfusion of adult mouse liver and in vitro mechanical stretching of human hepatic endothelial cells illustrate that mechanotransduction alone is sufficient to turn on angiocrine signals. When the endothelial cells are mechanically stretched, angiocrine signals trigger in vitro proliferation and survival of primary human hepatocytes. Our findings uncover a signalling pathway in vascular endothelial cells that translates blood perfusion and mechanotransduction into organ growth and maintenance.

We analysed liver development in mouse embryos and found that the liver grows substantially between embryonic day (E)12.5 and E13.5 (Fig. 1a–e). To assess the area of vascular perfusion, we injected fluorescently labelled *Ricinus communis* agglutinin into the hearts of mouse embryos that were cultivated ex vivo in whole embryo culture. Notably,

we revealed that liver perfusion starts at the liver periphery at E11.5 and extends towards the centre at E12.5, imminently before massive liver growth was observed (Fig. 1e–k). Further, 5-ethynyl-2'-deoxyuridine (EdU) incorporation and Ki-67 staining at E11.5 showed that hepatic cell proliferation was preferentially located in the perfused liver periphery (Extended Data Fig. 1), which suggests that vascular perfusion induces proliferation of hepatocytes and—in turn—growth of the developing liver.

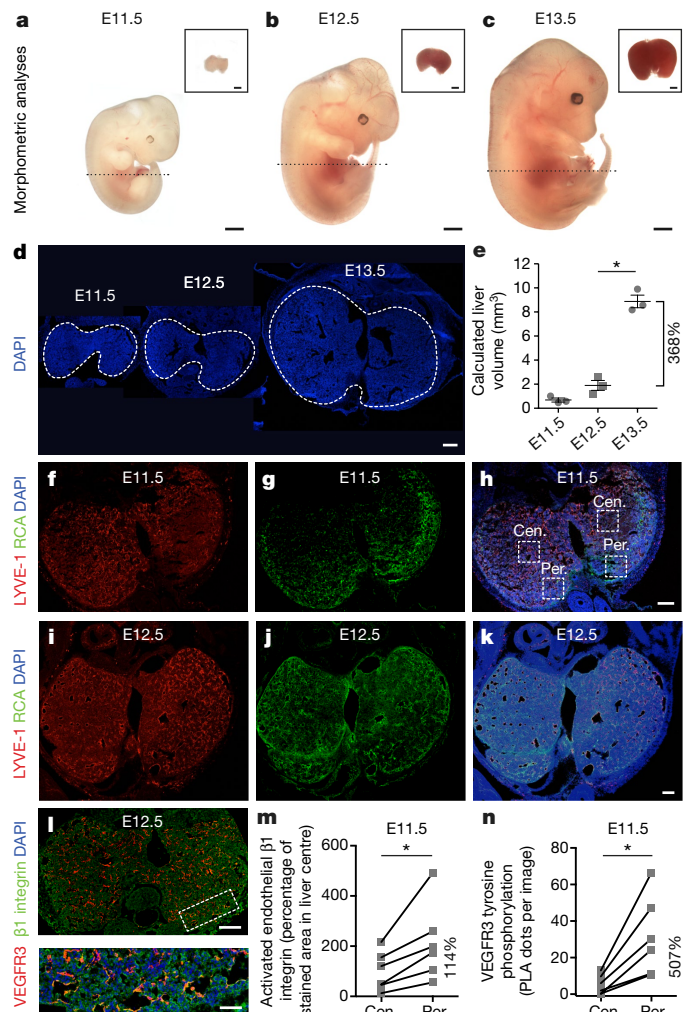


Fig. 1 | Vascular perfusion correlates with liver growth and activation of $\beta 1$ integrin and VEGFR3. **a–c**, Mouse embryos and their isolated livers (shown in insets, top right). **d**, Transverse sections (indicated by dotted black line in **a–c**) through embryonic livers (surrounded by white dashed lines), stained for cell nuclei (DAPI, blue). **e**, Calculated liver volumes ($n = 3$ embryos per stage, $*P = 0.0001$ (5.26; 8.71)). **f–k**, Transverse sections through livers stained for hepatic blood vessels (**f**, **i**) with LYVE-1 (red) and perfused vessels (**g**, **j**) with *R. communis* agglutinin (RCA, green). **h**, **k**, Merge with DAPI (blue). **l**, Top, hepatic blood vessels stained for $\beta 1$ integrin and VEGFR3. Bottom, magnification of liver region indicated in the top panel, with cell nuclei (DAPI, blue). **m**, **n**, Activation of endothelial $\beta 1$ integrin ($n = 6$ embryos, $*P = 0.0210$ (25.77; 202.20)) (**m**) and VEGFR3 tyrosine phosphorylation based on proximity ligation assay (PLA) ($n = 6$ embryos, $*P = 0.0118$ (8.88; 44.09)) (**n**) in liver centre (cen.) versus liver periphery (per.) as indicated by boxes in **h**. Scale bars, 1 mm (**a–c**), 500 μ m (insets in **a–c**), 300 μ m (**d**), 200 μ m (**f–k**, **l** top panel), and 50 μ m (**l** bottom panel). Data in **e** are mean \pm s.e.m. One-way ANOVA followed by Tukey's test (**e**) and two-tailed paired Student's *t*-tests (**m**, **n**), 95% confidence interval (lower confidence limit; upper confidence limit).

¹Institute of Metabolic Physiology, Heinrich-Heine University, Düsseldorf, Germany. ²Institute for Beta Cell Biology, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich-Heine University, Düsseldorf, Germany. ³German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany. ⁴Clinic for Gastroenterology, Hepatology and Infectious Diseases, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany. ⁵The Centre of Excellence in Translational Cancer Biology, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland. ⁶Department of Cardiovascular Medicine, Division of Cardiology, University Clinic Münster, Münster, Germany. ⁷European Institute for Molecular Imaging (EIMI), University of Münster, Münster, Germany. ⁸Institute for Clinical Diabetology, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich-Heine University, Düsseldorf, Germany. ⁹Division of Endocrinology and Diabetology, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany. ¹⁰Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich-Heine University, Düsseldorf, Germany. ¹¹These authors contributed equally: Linda Lorenz, Jennifer Axnick. *e-mail: lammert@hhu.de

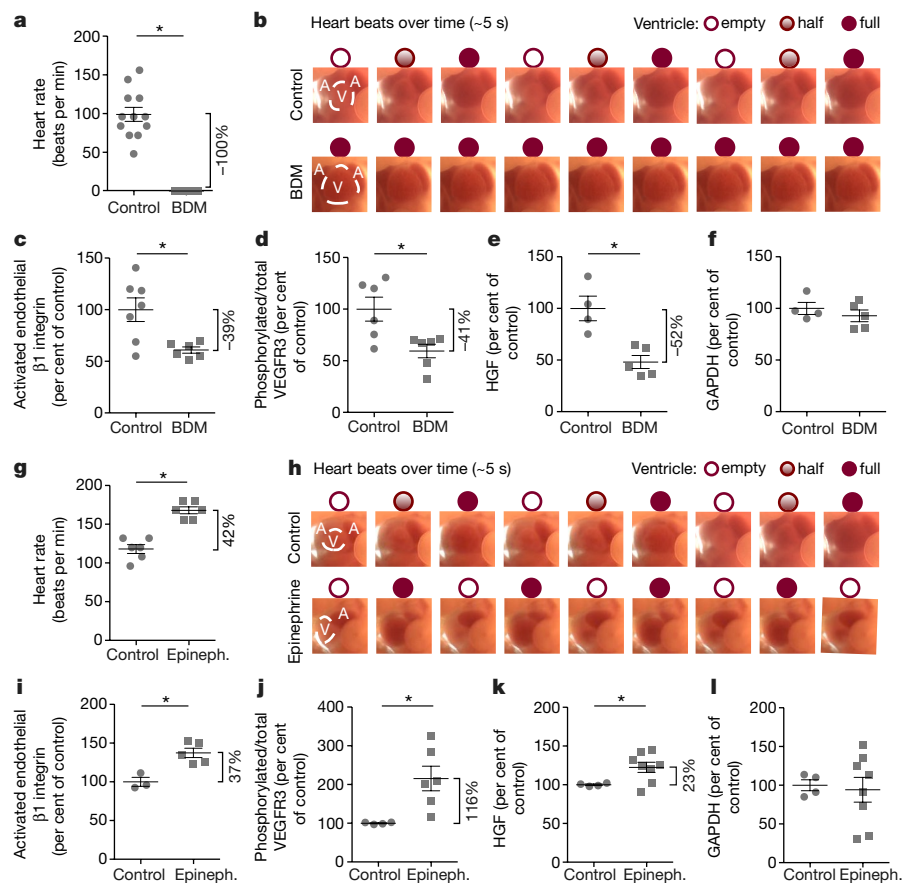


Fig. 2 | Heart rate correlates with activation of endothelial $\beta 1$ integrin, VEGFR3 and HGF production in whole embryo culture. **a–f**, Pharmacologic loss-of-perfusion experiments. **g–l**, Pharmacologic gain-of-perfusion experiments. **a, g**, Heart rates measured ex vivo: $n = 12$ control versus $n = 9$ E12.5 embryos treated with 2,3-butanedione monoxime (BDM) ($*P = 0.0001$ (79.19; 118.80)) (**a**) and $n = 6$ control versus $n = 6$ E12.5 embryos treated with epinephrine (epineph.) ($*P = 0.0001$ (33.78; 66.22)) (**g**). **b, h**, Time-lapse images of the beating heart of E11.5 embryos with schematics of ventricles (V, ventricle; A, atria). **c, i**, Quantification of activated endothelial $\beta 1$ integrin in embryonic livers: $n = 7$ control versus $n = 6$ BDM-treated E11.5 embryos ($*P = 0.0133$ (11.07; 67.00)) (**c**) and $n = 3$ control versus $n = 5$ epinephrine-treated E11.5 embryos ($*P = 0.0055$ (16.26; 58.43)) (**i**). **d, j**, Hepatic VEGFR3

tyrosine phosphorylation normalized to total VEGFR3: $n = 6$ control versus $n = 6$ BDM-treated E12.5 livers ($*P = 0.0165$ (9.70; 71.48)) (**d**) and $n = 4$ control versus $n = 6$ epinephrine-treated E12.5 livers ($*P = 0.0149$ (33.86; 197.10)) (**j**). **e, k**, Hepatic HGF protein concentration normalized to total protein: $n = 4$ control versus $n = 5$ BDM-treated pools of 5 E11.5 livers each ($*P = 0.0140$ (16.35; 87.45)) (**e**) and $n = 4$ control versus $n = 8$ epinephrine plus atropine-treated E12.5 livers ($*P = 0.0110$ (6.90; 38.12)) (**k**). **f, l**, Hepatic GAPDH protein concentration normalized to total protein: $n = 4$ control versus $n = 5$ BDM-treated pools of 5 E11.5 livers each ($P = 0.4112$ (−26.57; 12.28)) (**f**) and $n = 4$ control versus $n = 8$ epinephrine plus atropine-treated E12.5 livers ($P = 0.7497$ (−45.19; 33.69)) (**l**). Data are mean \pm s.e.m. Two-tailed unpaired Student's *t*-tests, 95% confidence interval (lower confidence limit; upper confidence limit).

Liver sinusoidal endothelial cells (ECs), which represent the largest population of ECs in the liver, are similar to lymphatic ECs in that both cell types express lymphatic vessel endothelial hyaluronan receptor-1 (LYVE-1) and VEGFR3^{6–8}. The latter is activated by mechanotransduction in the lymphatic endothelium, a process that requires the extracellular matrix receptor subunit $\beta 1$ integrin^{9–11}. Because these receptors co-localize on hepatic endothelium (Fig. 1l), we investigated whether the activation state of $\beta 1$ integrin and VEGFR3 corresponded to vascular perfusion level. Notably, more activated receptors were found in the perfused, growing liver periphery than the less-perfused non-proliferating liver centre (Fig. 1m, n and Extended Data Fig. 2a–d), while the total amount of endothelial $\beta 1$ integrin was not significantly different in these two areas (Extended Data Fig. 2e–g).

To analyse whether experimental manipulation of blood perfusion changes endothelial mechanotransduction in the developing liver, we designed loss- and gain-of-perfusion experiments in whole embryo culture (Fig. 2; Supplementary Videos 1, 2 are available online, see ‘Data availability’ in Methods). In the loss-of-perfusion study, the heartbeat of the mouse embryo was pharmacologically halted using 2,3-butanedione monoxime, and we found that endothelial $\beta 1$ integrin activation and VEGFR3 tyrosine phosphorylation were reduced (Fig. 2a–d; see Supplementary Fig. 1 for western blots). Conversely, in

the gain-of-perfusion study we observed more receptor activation in response to heart rate acceleration using epinephrine, with or without atropine (Fig. 2g–j). Decreasing or increasing the heartbeat frequency also lowered or enhanced, respectively, the protein concentration of hepatocyte growth factor (HGF) (Fig. 2e, k). By contrast, the concentration of glyceraldehyde-3-phosphate dehydrogenase (GAPDH), used as a housekeeping protein) was changed to a lesser extent (Fig. 2f, l). These data indicate that blood perfusion leads to mechanotransduction in hepatic endothelium, which in turn induces production of HGF, one of the key angiocrine signals required for liver growth and survival^{8,12–14}.

Next, we deleted the gene for $\beta 1$ integrin (*Itgb1*) in ECs to determine whether this previously described¹⁵ component of endothelial mechanosensing is required for angiocrine production of HGF during liver development (Fig. 3a–f). Notably, this deletion significantly lowered VEGFR3 tyrosine phosphorylation and protein content of HGF, but not GAPDH (Fig. 3d–f). Depletion of HGF was previously shown to result in embryonic lethality, reduced liver size and increased cell death in the liver periphery¹². A similar phenotype was observed upon EC-specific depletion of $\beta 1$ integrin, which also leads to smaller liver volumes (Fig. 3g–i), fewer EdU- and Ki-67-positive proliferating cells and more apoptotic cells in the liver periphery (Fig. 3j–l and Extended Data Fig. 3). Deletion of *Itgb1* using a different EC-specific

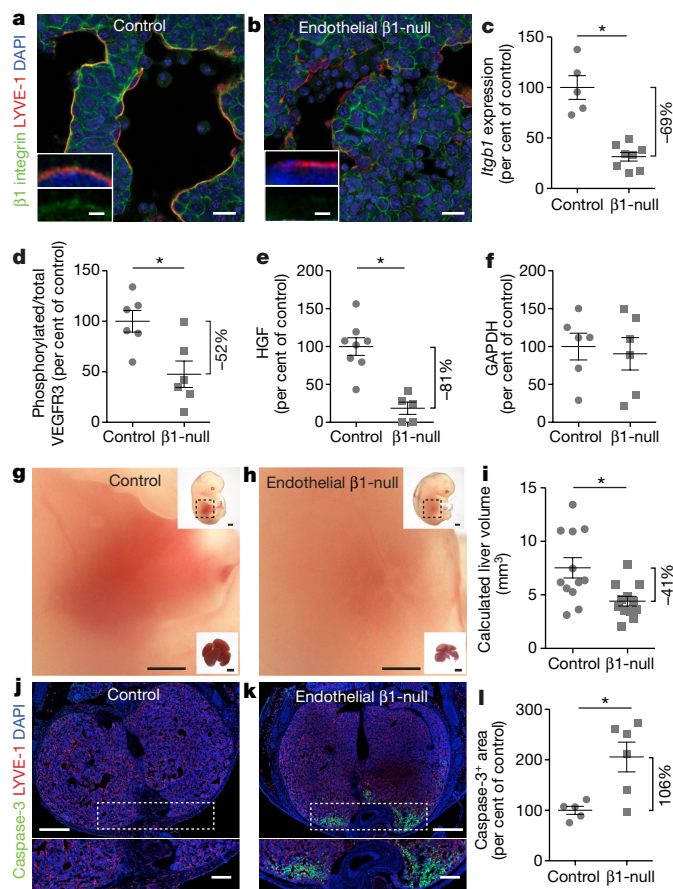


Fig. 3 | Endothelial $\beta 1$ integrin is required for VEGFR3 activation, HGF production, embryonic liver growth and survival. **a, b**, Transversal sections through E13.5 mouse livers with EC-specific heterozygous (control) and homozygous ($\beta 1$ integrin-null) depletion of $\beta 1$ integrin. Insets in bottom left of **a, b** show magnified images of control versus $\beta 1$ integrin-null mouse livers. **c–f**, Quantification of *Itgb1* mRNA in endothelial cells sorted from embryonic livers: $n = 5$ Cre-control versus $n = 8$ $\beta 1$ integrin-null ($\beta 1$ -null) livers ($*P = 0.0027$ (36.34; 100.70)) (**c**); VEGFR3 tyrosine phosphorylation normalized to total VEGFR3: $n = 6$ Cre-control and $\beta 1$ integrin-null livers each ($*P = 0.0117$ (14.60; 90.29)) (**d**); HGF protein: $n = 8$ control versus $n = 5$ $\beta 1$ integrin-null pools of two livers each ($*P = 0.0001$ (50.18; 112.50)) (**e**); and GAPDH protein: $n = 6$ Cre-control and $\beta 1$ integrin-null livers each ($P = 0.7397$ (–71.53; 52.58)) (**f**) in E12.5 liver lysates. **g, h**, Bright-field images of E13.5 mouse embryos (top right inset), their abdomen (as outlined on the embryo) (main panel) and isolated E14.5 livers (bottom right inset). **i**, Calculated E13.5 liver volumes ($n = 12$ livers per genotype, $*P = 0.0096$ (0.87; 5.35)). **j–l**, Transversal sections through E13.5 livers and quantification of cleaved caspase-3 positive areas ($n = 5$ control versus $n = 6$ $\beta 1$ integrin-null livers, $*P = 0.0143$ (30.32; 181.00)). Scale bars, 25 μm (**a, b**), 1 μm (insets in **a, b**), 500 μm (**g, h**, top panels in **j, k**, bottom right insets in **g, h**), 1 mm (top right insets in **g, h**), and 200 μm (bottom panels in **j, k**). Data are mean \pm s.e.m. Two-tailed unpaired Student's *t*-tests, 95% confidence interval (lower confidence limit; upper confidence limit).

Cre mouse line (*Cdh5-cre^{ERT2}* instead of *Flk1-cre^{ERT2}*; *Flk1* is also known as *Kdr*) also resulted in smaller livers, more apoptosis, lower levels of VEGFR3 activation and less HGF compared to controls (Extended Data Fig. 4). Because apoptotic cells were found next to blood vessels that were well-perfused (Extended Data Fig. 3g–i), we conclude that decreased angiocrine signalling rather than a lack of oxygen was responsible for the observed liver cell death.

To investigate whether VEGFR3 activation by vascular perfusion is dependent on endothelial $\beta 1$ integrin, gain-of-perfusion experiments were performed in *Itgb1*-deficient mouse embryos^{17,18}. In contrast to a heterozygous deletion of *Itgb1* in which activation of VEGFR3 was observed, depletion of $\beta 1$ integrin resulting from a homozygous

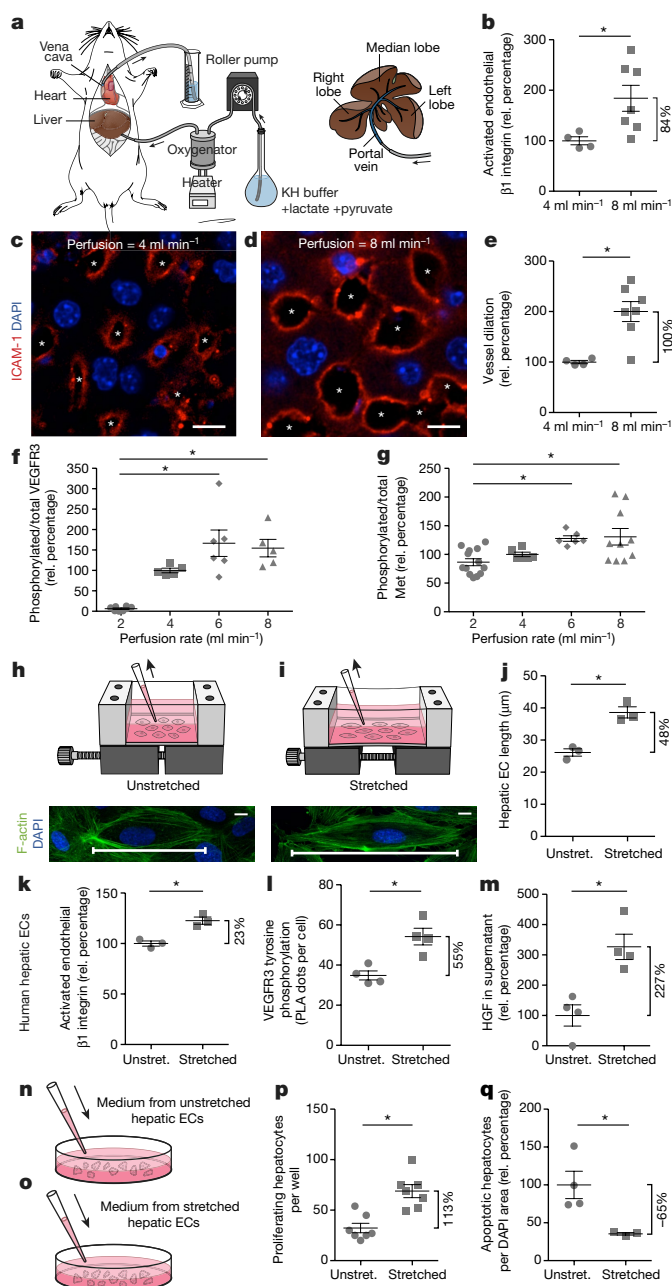
deletion of *Itgb1* was found to completely prevent the increased blood perfusion from activating VEGFR3 (Extended Data Fig. 5a, b). Further, tamoxifen-induced deletion of *Vegfr3* (also known as *Flt4*) significantly reduced liver size, increased the rate of apoptosis in the liver periphery and lowered the protein concentration of HGF compared to tamoxifen-injected controls (Extended Data Fig. 5c–i). The results show that during endothelial mechanotransduction, VEGFR3 is downstream of $\beta 1$ integrin and is also required for normal production of HGF.

To directly show that liver perfusion promotes angiocrine signalling even in the absence of the potential contributions of blood components, we perfused the adult mouse liver with a pre-warmed, oxygenated Krebs–Henseleit buffer (Fig. 4a). Notably, increasing the perfusion rate of the adult liver activated endothelial $\beta 1$ integrin and widened the vascular lumen of liver sinusoids (Fig. 4b–e). Similar vessel dilation was observed within the remaining right liver lobe after a two-thirds partial hepatectomy (Extended Data Fig. 6a–c), representing a strong inducer of adult liver growth¹⁹. Consistent with this vasodilation, contrast-enhanced ultrasound measurements indicated that—relative to the analysed liver area—more blood volume was present within the remaining liver lobe in each analysed animal, and that half of the animals also showed an increased blood flow velocity (Extended Data Fig. 6d–f). Further, tyrosine phosphorylation of VEGFR3 and Met (the receptor for HGF^{20,21}, also known as c-Met) increased at a rate proportional to ex vivo liver perfusion, within the range of 2 to 6 ml min^{–1} (Fig. 4f, g).

Next, we mechanically stretched primary human hepatic ECs for 90 min to mimic the mechanical stimulus induced by vascular lumen widening (Fig. 4h, i). The ECs were elongated upon stretching, and their $\beta 1$ integrin and VEGFR3 became activated (Fig. 4j–l and Extended Data Fig. 7a–d). In addition, mechanical stretching resulted in more interaction between $\beta 1$ integrin and VEGFR3 (Extended Data Fig. 7e–g), but did not significantly change tyrosine phosphorylation of VEGFR2 (encoded by *Kdr*) (Extended Data Fig. 7h–j). Then, we searched for angiocrine signals involved in liver growth and regeneration. Notably, more HGF was found in the supernatant of hepatic ECs upon mechanical stretching (Fig. 4m), and stretched hepatic ECs also released more interleukin 6 (IL-6) and tumour necrosis factor (TNF, also known as TNF α) (Extended Data Fig. 7k, l), two pro-inflammatory cytokines shown to be involved in liver regeneration^{22,23}. Higher activity of matrix metalloproteinase-9 (MMP9), which has previously been shown to activate HGF and to be required for normal liver regeneration^{24–27}, was also detected in the medium of mechanically stretched versus unstretched hepatic ECs (Extended Data Fig. 7m).

We then transferred the medium of human hepatic ECs onto in vitro cultured primary human hepatocytes, and observed a higher rate of proliferation and lower rate of apoptosis in the human hepatocytes treated with medium from stretched versus unstretched ECs (Fig. 4n–q, and Extended Data Fig. 8a–d). To investigate whether endothelial $\beta 1$ integrin activation is sufficient to promote hepatocyte proliferation, the hepatic ECs were treated with a $\beta 1$ integrin-activating antibody²⁸. This antibody significantly activated $\beta 1$ integrin on unstretched hepatic ECs (Extended Data Fig. 8e–g), and human hepatocytes exposed to the supernatant of antibody-treated ECs had increased proliferation compared to those exposed to the supernatant of untreated ECs (Extended Data Fig. 8h–j). Additionally, unconditioned medium containing this antibody or conditioned medium from ECs treated with an isotype control did not stimulate hepatocyte proliferation (Extended Data Fig. 8j).

Finally, we analysed data from 87 human glucose-tolerant volunteers (without diagnosis of diabetes mellitus) participating in the prospective, observational German Diabetes Study²⁹. More specifically, we looked for potential correlations between blood pressure and liver volume, which was determined by magnetic resonance imaging. Individuals were excluded when they were on antihypertensive medication or when no magnetic resonance imaging data were available (Extended Data Fig. 9a). Further, people who were obese (with a body mass index larger than 30 kg per m²) or diagnosed with hepatic steatosis (indicated by liver fat content larger than 5.5% relative to water) were analysed separately. Notably, we found a positive correlation between blood



pressure and liver volume in 42 individuals (Extended Data Fig. 9b, c), which was more pronounced when considering systolic blood pressure ($n = 42$, $r = 0.45$, $*P = 0.0022$ (0.17; 0.66); for all P values, numbers in parentheses indicate lower and upper confidence limits, respectively, of 95% confidence interval) compared to diastolic blood pressure. The correlation between systolic blood pressure and liver volume persisted when adjusted for liver fat content ($n = 42$, $r = 0.40$, $*P = 0.0098$ (0.10; 0.62)). By contrast, individuals with obesity or hepatic steatosis did not show any correlation between the systolic blood pressure and liver volume ($n = 9$, $r = -0.05$, $P = 0.9096$ (-0.69; 0.64)). Therefore, these data are consistent with the notion that mechanical forces in the blood vascular system are associated with the liver size in metabolically healthy individuals.

In sum, mechanical stretching of ECs during vasodilation induces angiocrine signals that support liver growth, survival and potentially contribute to liver regeneration (Extended Data Fig. 10; Supplementary Video 3 is available online, see 'Data availability' in Methods). An inductive, non-proliferating hepatic vasculature has previously been shown to contribute to hepatocyte proliferation in the first few days after partial hepatectomy⁸. According to the model, the decreased

Fig. 4 | Ex vivo liver perfusion and mechanical stimulation of hepatic ECs induce angiocrine signals. **a**, Schematics of an ex vivo perfusion experiment with a perfused mouse liver. KH, Krebs–Henseleit. **b**, Activation of endothelial $\beta 1$ integrin in ex vivo perfused livers: $n = 4$ livers at a flow rate of 4 ml min⁻¹ versus $n = 7$ livers at a flow rate of 8 ml min⁻¹ ($*P = 0.0164$ (20.74; 147.50)). Shown relative to activation at a flow rate of 4 ml min⁻¹. **c**, **d**, Sections from ex vivo perfused livers stained for blood vessels (asterisks). **e**, Vessel dilation: $n = 4$ livers at a flow rate of 4 ml min⁻¹ versus $n = 7$ livers at a flow rate of 8 ml min⁻¹ ($*P = 0.0021$ (51.67; 148.60)). Shown relative to dilation at a flow rate of 4 ml min⁻¹. **f**, Tyrosine phosphorylation of VEGFR3 normalized to total VEGFR3: $n = 8$ livers at a flow rate of 2 ml min⁻¹, $n = 5$ livers at a flow rate of 4 ml min⁻¹, $n = 6$ livers at a flow rate of 6 ml min⁻¹ and $n = 5$ livers at a flow rate of 8 ml min⁻¹ (2 versus 6 ml min⁻¹, $*P = 0.0001$ (91.32; 228.90); 2 versus 8 ml min⁻¹, $*P = 0.0001$ (75.64; 220.90)). Shown relative to ratio of phosphorylated to total VEGFR3 at a flow rate of 4 ml min⁻¹. **g**, Phosphorylated Met normalized to total Met: $n = 13$ livers at a flow rate of 2 ml min⁻¹, $n = 6$ livers at a flow rate of 4 ml min⁻¹, $n = 6$ livers at a flow rate of 6 ml min⁻¹ and $n = 10$ livers at a flow rate of 8 ml min⁻¹ (2 versus 6 ml min⁻¹, $*P = 0.0330$ (2.58; 79.93); 2 versus 8 ml min⁻¹, $*P = 0.0051$ (11.27; 77.19)). Shown relative to ratio of phosphorylated to total Met at a flow rate of 4 ml min⁻¹. **h**, **i**, Illustrations of how supernatants were taken from stretch chambers, with microscopic images of hepatic ECs shown in bottom panels. **j**, EC lengths ($n = 3$ stretch chambers each; $*P = 0.0061$ (6.34; 18.65)). **k**, **l**, Activation of endothelial $\beta 1$ integrin ($n = 3$ stretch chambers each, $*P = 0.0090$ (9.87; 35.31)) (**k**) and VEGFR3 ($n = 4$ stretch chambers each, $*P = 0.0112$ (6.84; 31.76)) (**l**). **m**, HGF protein concentrations in supernatant ($n = 4$ stretch chambers each, $*P = 0.0063$ (92.61; 360.60)). **n–q**, Illustrations of human hepatocytes supplemented with supernatant from hepatic ECs (**n**, **o**), with proliferation ($n = 7$ wells each; $*P = 0.0008$ (18.93; 54.22)) (**p**) and apoptosis of supernatant-treated hepatocytes ($n = 4$ and $n = 3$ wells treated with supernatant from unstretched and stretched hepatic ECs, respectively; $*P = 0.0368$ (7.40; 121.70)) (**q**). Scale bars, 10 μ m. Data are mean \pm s.e.m. Two-tailed unpaired Student's t -test (**b**, **e**, **j–q**), one-way ANOVA followed by Tukey's test (**f**, **g**), 95% confidence interval (lower confidence limit; upper confidence limit).

number of vessels in the remaining parts of the liver dilate to allow the same amount of blood to pass through the smaller amount of remaining liver tissue. This vasodilation results in circumferential EC stretching that induces the release of hepatocyte growth-promoting signals. After the growth of the hepatocytes in the first days after partial hepatectomy, the ECs start to proliferate⁸. Because small-calibre blood vessels are initially formed during angiogenesis, the ECs now experience increased shear stress. This change in mechanical stimulus might contribute to stunted liver growth via release of hepatocyte growth-inhibiting signals, as recently indicated³⁰. Our findings warrant further studies into the angiocrine responses of ECs to mechanical stimuli, particularly because many organs experience vasodilation—and thus EC stretching—when their physiological function is extensively needed and they start to grow.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0522-3>.

Received: 21 July 2016; Accepted: 16 August 2018;

Published online 26 September 2018.

- Cleaver, O. & Melton, D. A. Endothelial signaling during development. *Nat. Med.* **9**, 661–668 (2003).
- Rafii, S., Butler, J. M. & Ding, B. S. Angiocrine functions of organ-specific endothelial cells. *Nature* **529**, 316–325 (2016).
- Kostallari, E. & Shah, V. H. Angiocrine signaling in the hepatic sinusoids in health and disease. *Am. J. Physiol. Gastrointest. Liver Physiol.* **311**, G246–G251 (2016).
- Augustin, H. G. & Koh, G. Y. Organotypic vasculature: from descriptive heterogeneity to functional pathophysiology. *Science* **357**, eaal2379 (2017).
- Zaret, K. S. & Grompe, M. Generation and regeneration of cells of the liver and pancreas. *Science* **322**, 1490–1494 (2008).
- Si-Tayeb, K., Lemaigre, F. P. & Duncan, S. A. Organogenesis and development of the liver. *Dev. Cell* **18**, 175–189 (2010).

7. Mouta Carreira, C. et al. LYVE-1 is not restricted to the lymph vessels: expression in normal liver blood sinusoids and down-regulation in human liver cancer and cirrhosis. *Cancer Res.* **61**, 8079–8084 (2001).
8. Ding, B. S. et al. Inductive angiocrine signals from sinusoidal endothelium are required for liver regeneration. *Nature* **468**, 310–315 (2010).
9. Planas-Paz, L. et al. Mechanoinduction of lymph vessel expansion. *EMBO J.* **31**, 788–804 (2012).
10. Zhang, X., Groopman, J. E. & Wang, J. F. Extracellular matrix regulates endothelial functions through interaction of VEGFR-3 and integrin $\alpha_5\beta_1$. *J. Cell. Physiol.* **202**, 205–214 (2005).
11. Galvagni, F. et al. Endothelial cell adhesion to the extracellular matrix induces c-Src-dependent VEGFR-3 phosphorylation without the activation of the receptor intrinsic kinase activity. *Circ. Res.* **106**, 1839–1848 (2010).
12. Schmidt, C. et al. Scatter factor/hepatocyte growth factor is essential for liver development. *Nature* **373**, 699–702 (1995).
13. LeCouter, J. et al. Angiogenesis-independent endothelial protection of liver: role of VEGFR-1. *Science* **299**, 890–893 (2003).
14. Wang, L. et al. Liver sinusoidal endothelial cell progenitor cells promote liver regeneration in rats. *J. Clin. Invest.* **122**, 1567–1573 (2012).
15. Ingber, D. Integrins as mechanochemical transducers. *Curr. Opin. Cell Biol.* **3**, 841–848 (1991).
16. Benedito, R. et al. The Notch ligands Dll4 and Jagged1 have opposing effects on angiogenesis. *Cell* **137**, 1124–1135 (2009).
17. Licht, A. H., Raab, S., Hofmann, U. & Breier, G. Endothelium-specific Cre recombinase activity in flk-1-Cre transgenic mice. *Dev. Dyn.* **229**, 312–318 (2004).
18. Potocnik, A. J., Brakebusch, C. & Fässler, R. Fetal and adult hematopoietic stem cells require $\beta 1$ integrin function for colonizing fetal liver, spleen, and bone marrow. *Immunity* **12**, 653–663 (2000).
19. Forbes, S. J. & Newsome, P. N. Liver regeneration – mechanisms and models to clinical application. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 473–485 (2016).
20. Borowiak, M. et al. Met provides essential signals for liver regeneration. *Proc. Natl Acad. Sci. USA* **101**, 10608–10613 (2004).
21. Huh, C. G. et al. Hepatocyte growth factor/c-met signaling pathway is required for efficient liver regeneration and repair. *Proc. Natl Acad. Sci. USA* **101**, 4477–4482 (2004).
22. Michalopoulos, G. K. Liver regeneration after partial hepatectomy: critical analysis of mechanistic dilemmas. *Am. J. Pathol.* **176**, 2–13 (2010).
23. Böhm, F., Köhler, U. A., Speicher, T. & Werner, S. Regulation of liver regeneration by growth factors and cytokines. *EMBO Mol. Med.* **2**, 294–305 (2010).
24. Mohammed, F. F. et al. Metalloproteinase inhibitor TIMP-1 affects hepatocyte cell cycle via HGF activation in murine liver regeneration. *Hepatology* **41**, 857–867 (2005).
25. Kim, T. H., Mars, W. M., Stolz, D. B. & Michalopoulos, G. K. Expression and activation of pro-MMP-2 and pro-MMP-9 during rat liver regeneration. *Hepatology* **31**, 75–82 (2000).
26. Zhou, B. et al. Matrix metalloproteinases-9 deficiency impairs liver regeneration through epidermal growth factor receptor signaling in partial hepatectomy mice. *J. Surg. Res.* **197**, 201–209 (2015).
27. Olle, E. W. et al. Matrix metalloproteinase-9 is an important factor in hepatic regeneration after partial hepatectomy in mice. *Hepatology* **44**, 540–549 (2006).
28. Wayner, E. A., Gil, S. G., Murphy, G. F., Wilke, M. S. & Carter, W. G. Epiligrin, a component of epithelial basement membranes, is an adhesive ligand for alpha 3 beta 1 positive T lymphocytes. *J. Cell Biol.* **121**, 1141–1152 (1993).
29. Szendroedi, J. et al. Cohort profile: the German Diabetes Study (GDS). *Cardiovasc. Diabetol.* **15**, 59 (2016).
30. Manavski, Y. et al. Endothelial transcription factor KLF2 negatively regulates liver regeneration via induction of activin A. *Proc. Natl Acad. Sci. USA* **114**, 3993–3998 (2017).

Acknowledgements This work was supported by Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Centres SFB 974 ('Communication and Systems Relevance during Liver Damage and Regeneration'), SFB 1116 ('Master switches in cardiac ischemia'), IRTG 1902 ('Intra- and interorgan communication of the cardiovascular system') as well as DFG LA1216/6-1 ('Investigation of the role of vascular endothelium in blood glucose metabolism'), the German Center for Diabetes Research (DZD e.V.), the Federal Ministry of Health, the Ministry of Culture and Science of North Rhine-Westphalia, the Academy of Finland, the Novo Nordisk Foundation and Helsinki Institute of Life Science (HiLIFE). We are also grateful to Y. Koh for schematic illustrations, M. Astrachan and his team (XVIVO scientific animation) for the animation, L. S. Hilger for establishing VEGFR3 western blots, S. Jakob, B. Bartosinska, A. Köster and T. Zobel (Centre for Advanced Imaging) for technical support as well as B.-F. Belgardt, C. Bridges, M. Kelly-Goss and M. Gearing for critical reading of the manuscript.

Reviewer information *Nature* thanks G. Michalopoulos, E. Tzima and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.L. and J.A. performed the majority of the experiments; J.A. performed most experiments on mouse embryos—many of which were repeated and/or finalized by L.L., C.H. and S.U.—and contributed to correlation analyses; L.L. performed the experiments on human cells and the analyses of ex vivo perfused livers; T.B. performed the first embryonic experiments, planned and designed the first ex vivo liver perfusion experiments along with N.E. and performed partial hepatectomy experiments; C.H. contributed to manipulation of mouse embryos and whole embryo culture. S.F. and H.N. were supervised by K.A., and provided VEGFR3 knockout embryos, genotyping and knockout efficiencies. D.H. planned, designed and supervised the ex vivo liver perfusion experiments performed by N.E. R.H. and J.S. performed and analysed the contrast-enhanced ultrasound measurements. K.M., K.B. and M.R. recruited, screened and phenotyped human individuals within the German Diabetes Study of which M.R. is the principal investigator; J.-H.H. and K.B. performed the magnetic resonance imaging and ^1H magnetic resonance spectroscopy analyses along with M.R.; and O.K. performed the correlation analyses and adjustments. D.E. isolated hepatic ECs from mouse embryos and was involved in immunohistochemical analyses of mouse embryos, discussions, data management and statistical evaluations. E.L. supervised and guided J.A., T.B., C.H., S.U. and L.L. during their experiments and wrote the manuscript with the help from J.A. and L.L. All authors read and contributed to the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0522-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0522-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse strains. C57BL/6J (Janvier) mice were used for wild-type studies and whole embryo culture (WEC). For genetic deletion of *Itgb1*, *Flk1-cre*, *Cdh5-cre^{ERT2}* and *Itgb1-loxP* mice were used^{9,16–18}. To genetically delete *Vegfr3*, *Rosa26-cre^{ERT2}* and *Vegfr3-loxP* mice were used^{31,32}. All mice were on C57BL/6J background. For induction of Cre-mediated gene recombination in embryos, the pregnant mother received either two consecutive intraperitoneal injections of 100 μ l tamoxifen (Sigma, 18 mg/ml dissolved in peanut oil) at E9.5 and E10.5 for *Itgb1* deletion or two consecutive intragastric doses of 75 μ l 4-OH-tamoxifen (Sigma, 25 mg/ml dissolved in ethanol and olive oil) by gavage at E10.5 and E11.5 for *Vegfr3* deletion. All experiments were performed according to the German (Animal Ethics Committee of the Landesamt für Natur, Umwelt und Verbraucherschutz, North Rhine–Westphalia) and the Finnish (Committee for Animal Experiments of the District of Southern Finland) animal protection laws.

Mouse embryonic experiments. Experiments with mouse embryos—in particular, the WEC experiments—were performed as previously described³³. In brief, embryos were staged by the day of the presence of a vaginal plug (E0.5). For injection into the heart or the liver, an InjectMan NI 2 microinjector connected to a FemtoJet pump (Eppendorf) was used. Borosilicate glass capillaries with filament (Harvard Apparatus) were used for injections; the capillaries were pulled with a needle puller (Narishige). One per cent Fast Green (AppliChem) was added to all injected substances to visualize the injection site. Directly after isolation or following the injection, embryos were cultured in a WEC system (RKI Ikemoto Scientific Technology) in 1,600 μ l DMEM (Gibco) and 400 μ l fetal bovine serum (Gibco) per WEC bottle at 37°C, oxygenated with a flow rate of 75 cc of 5% CO₂ in 95% O₂, and rotated at 25 rotations per minute. Following embryo isolation or after WEC, embryos were fixed with 4% paraformaldehyde (PFA) over night; for protein analyses, livers were isolated in ice-cold PBS with phosphatase and protease inhibitors (PhosStop and cComplete Protease Inhibitor Cocktail, Sigma).

Vascular painting and EdU incorporation. For vascular painting, fluorescently labelled (FITC) RCA (10% diluted in PBS, Sigma) was injected into the beating heart of E11.5 and E12.5 embryos. Afterwards, the embryos were cultured in WEC for 3 h to allow the RCA to bind (paint) all perfused blood vessels. After WEC, the embryos were fixed and processed for liver isolation and immunofluorescent analyses. For detection of proliferating cells, EdU (thymidine base analogue, Life Technologies) was diluted to 1 mg/ml and injected into the middle of each lobe of the embryonic liver until the whole lobe was coloured with injection dye. EdU was allowed to incorporate into all proliferating cells for 3 h. After fixation and cryo-sectioning of the embryos, the EdU Click-iT assay was performed according to the protocol provided by the company (Thermo Fisher).

Loss-of-perfusion and gain-of-perfusion experiments. For loss- and gain-of-perfusion experiments, E11.5 and E12.5 wild-type mouse embryos or mouse embryos with or without genetic deletion of *Itgb1* were isolated and cultured in WEC for 15 min to 1 h. For loss-of-perfusion experiments, BDM (Sigma) diluted in PBS was added at a final concentration of 50 mM to the culture medium to arrest the beating heart. The success of heart block was controlled in the first minutes of WEC. For gain-of-perfusion experiments, final concentrations of 2.5 mg/ml or 5 mg/ml (–)–epinephrine-(+)-bitartrate salt (Sigma) diluted in H₂O or 5 mg/ml epinephrine plus 1 μ M atropine sulfate salt monohydrate (Sigma) both diluted in H₂O were added to the culture medium. The success of increasing the heart rate was controlled in the first minutes of WEC and after cultivation. The heart rate was measured on a heating plate at 37°C using a stereomicroscope with digital camera (Nikon, SMZ1500). Supplementary Videos 1, 2 (see ‘Data availability’) were made using the same microscope, but without any heating plate. For quantification of heart rate, VEGFR3 phosphorylation, HGF and GAPDH protein concentrations, epinephrine-treated embryos with a heart rate higher 140 beats per minute and control embryos with a heart rate lower 140 beats per minute were used. Embryos were fixed and prepared for immunohistochemistry, or their livers were isolated on ice for protein analyses.

Ex vivo perfusion of adult mouse livers. This method is modified after a previously described method used for rats³⁴. For open, non-recirculating perfusion, 12–14-week-old male C57BL/6J mice (Janvier) were killed either by cervical dislocation or by blood replacement with 37°C preheated bicarbonate-buffered Krebs–Henseleit saline solution containing pyruvate (0.3 mM) and L-lactate (2.1 mM), after narcotization with a mixture of xylazine (Bayer) and ketamine (Pfizer). The portal vein was cannulated with a 20G intravenous catheter (Hospira VENISYSTEMS, Abbocath-T) connected to a perfusion system. For the efflux, the inferior vena cava was also cannulated with a 20G intravenous catheter (Hospira VENISYSTEMS, Abbocath-T), which drained the perfusate into a measuring cylinder to monitor flow rates. The Krebs–Henseleit saline solution, containing phosphatase inhibitors (PhosSTOP, Sigma) for protein analyses, was pumped by a peristaltic pump (at different flow rates) through an oxygenator (oxygenation with carbogen gas, 95% O₂ and 5% CO₂), a heater, a bubble trap (to prevent liver emboli) and, finally, through the cannulated liver. After 15 min of perfusion, the liver was

deep-frozen in liquid nitrogen for protein analyses and after 1 h of perfusion, the liver was perfused with 4% PFA for fixation followed by immunohistochemical analyses.

Primary human cell culture experiments. Human hepatic ECs—sold as ‘liver sinusoidal microvascular endothelial cells’ (donors: female, 59 years old, Caucasian, body mass index = 18 kg/m², cause of death: anoxia secondary to cardiovascular disease; and male, 52 years old, Caucasian, body mass index = 30.6 kg/m², cause of death: anoxia)—were purchased from PELOBiotech; the hepatic ECs were isolated from collagenase type 1-digested peripheral liver tissue via sequential cell sorting with anti-CD146, anti-CD31 and anti-VEGFR2 antibodies. More than 95% of the isolated cells were reported by the manufacturer to have cytoplasmic immunofluorescent staining for von Willebrand factor, PECAM1 and Di-I-Ac-LDL uptake. In addition, the identity of these hepatic ECs was independently confirmed using gene expression analyses with appropriate controls and immunocytochemistry for von Willebrand factor. The ECs and cryo-preserved human hepatocytes were cultured in a humidified atmosphere at 5% CO₂ and 37°C. ECs were grown in microvascular endothelial cell growth medium kit enhanced (PELOBiotech), plated on stretch chambers (STREX) or on 12- or 24-well plates, both pre-coated with speed coat solution (PELOBiotech), and used up to passage 5. Hepatic ECs on stretch chambers were mechanically stretched for 1.5 h in medium without supplements, either with an automated cell-stretching system (STREX, STB-140) or with a manual cell-stretching system (STREX, STB-10). ECs on 12-well plates were stimulated for 1.5 h with 1 μ g/ml β 1 integrin activating antibody (R&D, MAB17782) or with 1 μ g/ml mouse IgG₁ isotype control antibody (R&D, MAB002). Frozen hepatocytes were suspended in cryo-preserved hepatocyte recovery medium and subsequently cultured in Williams E medium plus thawing/plating supplement pack, followed by cultivation (without expansion) in Williams E medium plus cell maintenance supplement pack (Thermo Fisher). For co-culture experiments, conditioned medium of mechanically unstretched versus stretched ECs or of β 1 integrin-stimulated ECs was transferred to human hepatocytes plated on Permanox Chamber Slides (ThermoFisher), pre-coated with speed coat solution plus 50 μ g/ml rat tail collagen I (ThermoFisher). Hepatocytes were incubated for 6 h (a time found to be optimal for co-culturing) with conditioned medium, plus 1 mg/ml EdU to allow immunohistochemical analyses of cell proliferation.

Partial hepatectomy in mice. To induce liver regeneration, a two-thirds partial hepatectomy was performed³⁵. Mice were anaesthetized with isoflurane (2% in 2 l per min O₂), and 5 mg/kg carprofen was injected intraperitoneally as an analgesic agent. The median and left liver lobes were gently moved aside with a PBS-moistened cotton tip to cut the falciform ligament and the membrane between the caudate and left liver lobe. The left and median lobes were ligated with a silk thread (Mersilene, Ethicon). Sufficiently ligated lobes changed their colour and could be cut above the suture by using micro-scissors. After potential bleedings were staunched, the abdominal cavity was flushed with a saline solution. After removing the saline with medical swabs, the peritoneum was closed with a suture (Vicryl Plus, Ethicon). The skin was closed with wound clips (FST) and wiped with disinfectant. Sham-operated mice underwent the same procedure without ligation of the lobes. For tissue fixation, blood was replaced with PBS at a physiologic perfusion rate over the left heart ventricle using a peristaltic pump (ISMATEK, ISM827B). For immunohistochemical analyses, 4% PFA was subsequently perfused into the circulation before liver removal.

Contrast-enhanced ultrasound measurements in mice. Contrast-enhanced ultrasound measurements were performed on a Vevo2100 ultrasound system (FUJIFILM Visualsonics) at 18 MHz, equipped with a MS-250 transducer operating in a contrast specific imaging mode. A single bolus of 30 μ l Vevo Micromarker Non-Targeted Contrast Agent (FUJIFILM Visualsonics) was injected intravenously via a tail-vein catheter before each measurement, using an automated Vevo Infusion Pump (FUJIFILM Visualsonics). Liver mass was quantified using VevoCQ Software (FUJIFILM Visualsonics). To ensure that always the same area was examined, liver was measured just above the kidney. Anatomically, the right liver lobe encloses the anterior tip of the right kidney and this lobe was not affected by the partial hepatectomy surgery. Measurements were based on the re-perfusion of the contrast agent after the signal had been destroyed with a high-power ultrasound signal (burst). Peak enhancement of the contrast agent (in linear arbitrary units) was used for calculation of relative blood volume, whereas time-to-peak of contrast agent enhancement was used to calculate relative blood flow velocity.

Cryo-sectioning of embryos and liver tissue. After WEC or directly following their isolation, embryos or livers were fixed with 4% PFA overnight at 4°C and subsequently equilibrated in 15% and 30% sucrose for cryo-preservation. Afterwards, embryos or embryonic livers were placed transversally—and the left or right lobe of adult livers with the visceral side down—into Peel-A-Way embedding molds (Polysciences), filled with OCT compound embedding medium (Tissue-Tek, Sakura Finetek GmbH), frozen down and stored at –80°C. A cryostat microtome HM 560 (Thermo Fisher Scientific) and MX35 premier microtome blades (Thermo Fisher

Scientific) were used to obtain consecutive 12- μ m cryo-sections that were placed onto SuperFrost-slides (Thermo Fisher Scientific).

Immunostaining, EdU and cell death staining, and morphometric analyses. For immunostaining of cryo-sections and primary human cells, the following primary antibodies were used: goat anti-mouse LYVE-1 (R&D, AF2125), rat anti- β 1 integrin (Millipore, MAB1997), rat anti-activated β 1 integrin (BD Bioscience, 553715), rabbit anti-Ki-67 (Sigma, AB9260), rabbit anti-caspase3 active (Sigma, C8487), rat anti-PECAM-1/CD31 (BD Bioscience, 550274), goat anti-ICAM-1 (R&D, AF796), goat anti-VEGFR3 (R&D, AF743), rabbit anti-HNF4 α (Cell Signaling, C11F12), mouse anti-human activated β 1 integrin (Merck Millipore, MAB2079Z). Secondary antibodies were as follows: donkey anti-rabbit/goat AF555 (Molecular Probes, A31572/ A2143), donkey anti-rabbit/goat/rat AF488 (Molecular Probes, A21206/ A11055/ A21208), donkey anti-rat/goat/rabbit/mouse Cy5 (Jackson ImmunoResearch, 712-175-153/705-175-147/711-175-152/715-175-150), AlexaFluor488 phalloidin (Thermo Fisher, A12379), DAPI (Sigma, D9542-1MG). EdU Click-iT reaction (Life Technologies) was performed on cryo-sections of WEC embryos or on Permax chamber slides with primary human hepatocytes after treatment with different media. Apoptosis in human hepatocytes was determined using the in situ cell death detection kit (TUNEL), TMR red, according to the manufacturer's instructions (Roche, 12156792910).

For morphometric analyses, the DAPI-stained liver area was determined by circulation with the marker tool in FIJI (ImageJ, NIH). For E11.5 embryos every 8th liver section, for E12.5 embryos every 10th liver section, and for E13.5 embryos every 12th liver section was imaged. For calculation of the liver volume, the liver areas of all imaged sections were extrapolated by multiplying the area of the liver with the number of tissue sections and section thickness. For quantification of the vascular lumen in liver sinusoids, the inner vessel lumen was encircled using the marker tool in FIJI (ImageJ, NIH).

Proximity ligation assays. PLA was used to detect areas of tyrosine phosphorylation close to VEGFR3 or VEGFR2, or to detect co-localization of VEGFR3 and β 1 integrin. The PLA was performed on cryo-sections and stretching chambers according to the protocol recommended by the company (Olink Bioscience). For detection of tyrosine phosphorylation, a mouse anti-phosphotyrosine antibody (Millipore, 05-1050) was used in combination with either goat anti-VEGFR3 (R&D, AF743 and AF349) or goat anti-VEGFR2 (R&D, AF357). For counter-staining, rabbit anti-mouse LYVE-1 (Abcam, AB14917) or rabbit anti-mouse ICAM-1 (proteintech, 10020-1-AP) was used. To determine colocalization of VEGFR3 and β 1 integrin, mouse anti- β 1 integrin (Chemicon, MAB1987) and goat anti-VEGFR3 (R&D, AF349) antibodies were used. For PLA labelling, anti goat-Plus-Probe (Sigma, Duolink DUO92003), anti mouse-Minus-Probe (Sigma, Duolink DUO92004), and Duolink in situ detection reagent orange (Sigma, Duolink DUO92007) were used.

Imaging and image analysis. Bright-field images were taken with a Nikon SMZ1500 stereomicroscope. For imaging all immunofluorescent sections, an Axio Vert LSM710 confocal laser scanning microscope (Zeiss) and Leica SP8 (Leica) were used. If images were edited, images that were compared with each other were handled in the same way. For area quantification, a threshold was set for the images to be analysed. For the analysis of endothelial cell staining, a mask was used to only measure the staining in the endothelial cell area (the latter stained for LYVE-1). Proliferating cells and PLA dots were counted either manually or using a counting macro in FIJI (ImageJ).

Magnetic-activated cell sorting, reverse transcriptase PCR and real-time PCR. Magnetic-activated cell sorting was used to isolate ECs from embryonic livers. Livers were cut into small pieces, dissociated using the gentleMACS Dissociator (Miltenyi Biotec) and digested using collagenase type 2 (Worthington). Anti-CD146 mouse MicroBeads (Miltenyi Biotec, 130-092-007) were used to isolate hepatic ECs from the digested embryonic livers in a magnetic field (QuadroMACS Separators, Miltenyi Biotec). For gene expression analysis, RNA was isolated using the RNeasy Mini Kit (Qiagen) and cDNA was synthesized with M-MLV Reverse Transcriptase (Promega), according to the manufacturer's instructions. *Itgb1* mRNA expression was quantified by qPCR using the Mx3000P qPCR System (Agilent Genomics) and the Brilliant III Sybr Green qPCR Mastermix (Agilent). The following primer sequences were used for amplification: *Hprt*, GCTGGTGAAAAGGACCTCT and CACAGGACTAGAACACCTGC; *Itgb1*, ATGCCAAATCTTGCGGAGAAT and TTTGCTGCGATTGCTGACATT. Relative gene expression of *Itgb1* in relation to *Hprt* was calculated according to a previously published method³⁶.

Enzyme-linked immunosorbent assays, MMP9 assay and western blot. Embryonic livers were isolated in ice-cold PBS with phosphatase and protease inhibitors. Afterwards, they were lysed in 75 μ l (E12.5) or 100 μ l (E11.5) ice-cold lysis buffer and sheared with a 1-ml syringe (Terumo). For E11.5 embryos, five livers were pooled into one lysate to get a sufficient amount of protein. E12.5 and E13.5 livers were lysed separately. Parts of deep-frozen adult livers (approximately 20 mg) were placed with 500 μ l lysis buffer into a gentleMACS M Tube (Miltenyi

Biotec) for tissue homogenization on gentleMACS Dissociator (Miltenyi Biotec) using the protein dissociation program. Lysis buffer contained HEPES 50 mM (Sigma), NaCl 150 mM (Sigma), Glycerol 10% (Sigma), Triton X-100 1% (Sigma), PhosSTOP phosphatase inhibitor (Sigma) and Complete cocktail protease inhibitors (Sigma). After homogenization, lysates were centrifuged at 13,000g for 10 min at 4 °C to remove cell trash, measured for protein content using BCA assay (Pierce BCA Protein Assay Kit, Thermo Fisher Scientific) and stored at -80 °C. Supernatants of human hepatic ECs were centrifuged at 13,000g for 10 min at 4 °C and immediately used for enzyme-linked immunosorbent assays (ELISAs). All ELISAs were performed according to the protocols provided by the company (R&D). DuoSet phospho-c-Met was normalized to total-c-Met (R&D, DY2480, DY2358). Mouse HGF DuoSet ELISA (R&D, DY2207) or mouse/rat HGF Quantikine ELISA Kit (R&D, MHG00) were used for mouse liver lysates. GAPDH was used as control protein during ELISA experiments (R&D, DY25718). Human HGF, IL-6 and TNF- α DuoSet ELISAs (R&D, DY294, DY206, DY210) were used for cell culture supernatants. To determine MMP9 activity in supernatants of human hepatic ECs, the supernatants were centrifuged at 13,000g for 10 min at 4 °C and immediately used for human active MMP-9 fluorokine assay following the manufacturer's instructions (R&D, F9M00). For western blotting, lysed samples were prepared with 2 \times Laemmli sample buffer (Bio-Rad) plus 2-mercaptoethanol (Carl Roth) and heated at 95 °C for 5 min. The same amounts of protein were loaded on Mini-Protein TGX Stain-Free Protein Gels (Bio-Rad) and transferred with the Trans-Blot Turbo Transfer System (Bio-Rad). Membranes were blocked with PBS containing 5% bovine serum albumin (Appligene) supplemented with 0.5% Tween (Sigma), and incubated with primary antibodies over night at 4 °C; these were rabbit anti-VEGFR3 (Signallway Antibody, 21410), rabbit anti-phospho-VEGFR3 (Cell Applications, CY1115), and rabbit anti-GAPDH (Abcam, Ab9484). After washing, membranes were incubated with goat anti-rabbit IgG, HRP-linked antibody (Cell Signaling, 7074S) for 45 min at room temperature. For protein detection Clarity Western ECL Substrate (Bio-Rad) and ChemiDoc MP Imaging System (Bio-Rad) were used. Western blots were densitometrically analysed using Image Laboratory Software Version 5.2 (Bio-Rad).

Studies of liver volume and blood pressure in humans. Data from 87 volunteers (without diagnosis of diabetes mellitus) from the German Diabetes Study were analysed (NCT01055093; approved by the ethics committee of the Medical Faculty at Heinrich-Heine University; reference numbers 2478 and 4508). Written informed consent was obtained from all individuals participating in the study. Blood pressure was measured between 08:45 and 13:30 using the automatic 705IT OMRON equipment (OMRON Healthcare GmbH) according to the guideline 93/42/EEC and Euronorm EN1060²⁹. Human individuals were in sitting position, and blood pressure was measured once on the left arm and thrice on the right arm. Blood pressure values were given as the mean of the last two out of three measurements obtained from the right arm. Magnetic resonance imaging (MRI) measurements to assess liver volumes were performed in overnight-fasted volunteers between 07:00–10:30 on a 3.0-T MR scanner with a body coil (Achieva X-series, Philips Healthcare). MRI was acquired using transverse multi-slice turbo spin echo sequences (recovery time (TR)/echo time (TE)_{eff} = 505/38 ms) with a turbo factor of 7. The field of view of MRI was 50.0 \times 46.9 cm² with the acquisition matrix size of 252 \times 189, and resulted MRI were reconstructed with the matrix size of 256 \times 256. Liver volume was quantified using the manual segmentation software available on the Philips MR console, as reported previously³⁷. To measure liver fat content by ¹H magnetic resonance spectroscopy (MRS), the location of liver was first identified using a quick localizer. The volume of interest was 25 \times 25 \times 25 mm³, which was carefully placed in the posterior region of the liver, avoiding major vessels and gall bladder. ¹H MRS was acquired using stimulated echo acquisition mode sequence (TR, 4,000 ms; TE, 10 ms; mixing time(TM), 16 ms). Both water-suppressed and non-suppressed ¹H MRS of the liver were taken in the identical voxel, and non-water-suppressed spectra served as an internal reference for fat quantification, as previously reported³⁷. All liver spectra were processed using jMRUI (Consortium of the EU project 'Advanced Signal Processing for Medical Magnetic Resonance Imaging and Spectroscopy'). The absolute percentage of fat content was determined as the ratio of intensities of the methylene (-CH₂)_n- peak in triglycerides at 1.3 ppm to the entire signal intensities of water and methylene fat signals (100 \times fat/(water + fat)).

Statistics and reproducibility. Except for the clinical data (that are mean \pm standard deviation) and paired data, all data are shown as mean \pm s.e.m. Two-tailed Student's *t*-tests (paired or unpaired with Welch correction) and one-way ANOVA with post-hoc tests were performed, if not mentioned otherwise, for statistical analyses using Excel (Microsoft) or PRISM 6 (Graph-Pad) software. *P* < 0.05 was considered statistically significant, and *P* values below 0.0001 are stated as *P* = 0.0001. Outliers were excluded after performing a Grubbs' test for outliers (*P* = 0.05) using PRISM 6 for *n* \geq 4 unpaired samples or data. Ninety-five per cent confidence intervals with lower and upper confidence limits are given for each comparison. Correlations are Pearson correlation coefficients with confidence

intervals and *P* values calculated after Fisher's *z*-transformation. Adjusted correlations or partial correlation coefficients are standard Pearson correlation coefficients between the residuals of two initial variables after regression on the adjusting variables. For correlation analyses, SAS software version 9.4 (SAS Institute) was used.

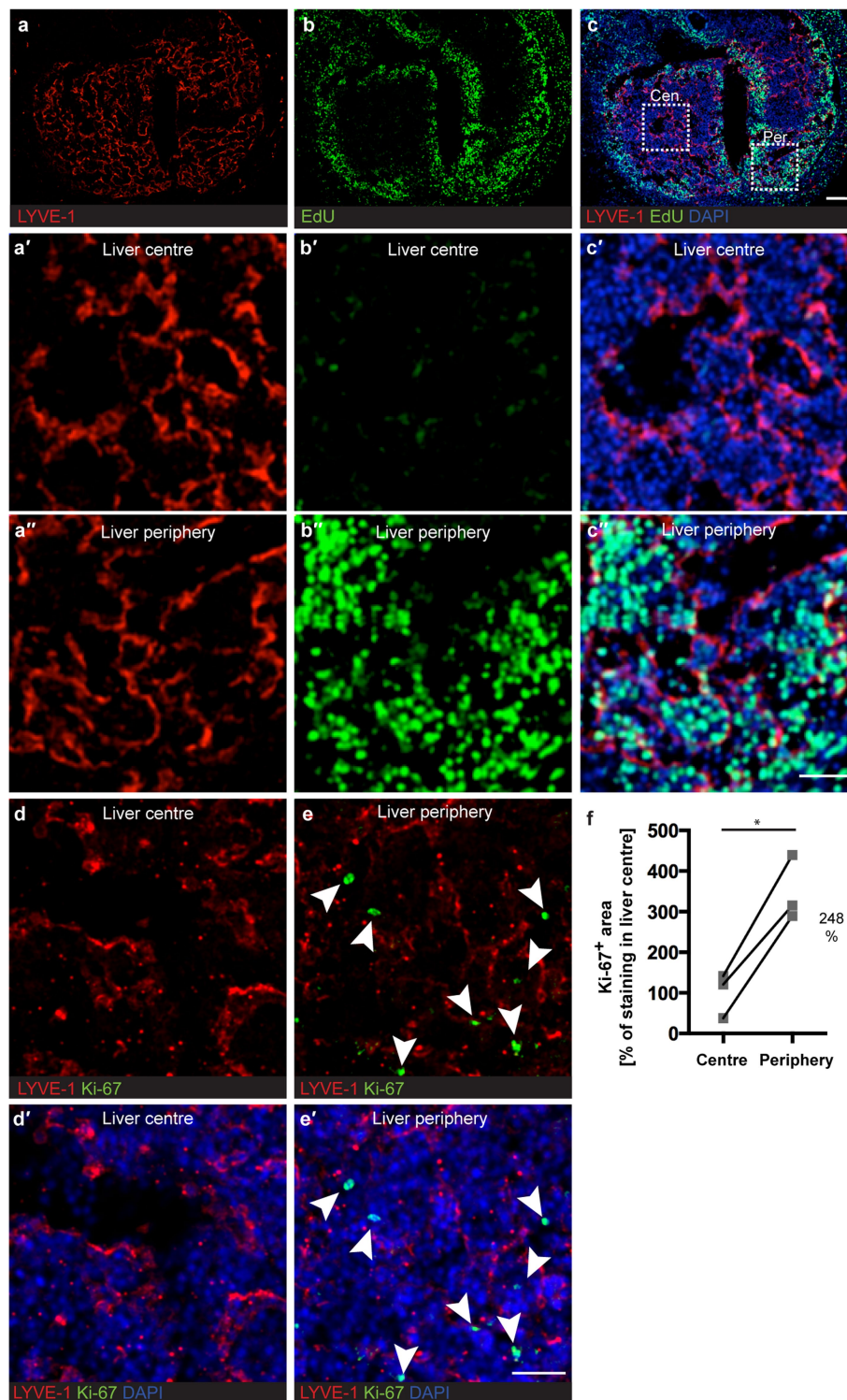
The following experiments were replicated: vascular painting of mouse embryos (shown in Fig. 1f–k); heart rate measurements (shown in Fig. 2a, g); GAPDH ELISA measurements (shown in Fig. 2l); β 1 integrin and LYVE-1 immunostaining (shown in Fig. 3a, b); phosphorylation of VEGFR3 in ex vivo perfused livers (shown in Fig. 4f); quantification of VEGFR3 tyrosine phosphorylation in stretched versus unstretched human hepatic ECs (shown in Fig. 4l); quantification of HGF protein concentration in supernatants from stretched versus unstretched ECs (shown in Fig. 4m); quantification of human hepatocyte proliferation after incubation of hepatocytes with medium from stretched versus unstretched ECs (shown in Fig. 4p) using hepatocytes from a different human donor; EdU and LYVE-1 staining of mouse embryos (shown in Extended Data Fig. 1a–c); RCA, caspase-3 and LYVE-1 staining of *Itgb1*-deficient embryos (as shown in Extended Data Fig. 3h, i); quantification of IL-6 and MMP9 activity in supernatants from stretched versus unstretched human ECs (shown in Extended Data Fig. 7k, m); quantification of β 1 integrin activation in human ECs treated with an activating β 1 integrin antibody (shown in Extended Data Fig. 8g). The other experiments, such as the correlation analysis with the German Diabetes Study cohort, were not replicated.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

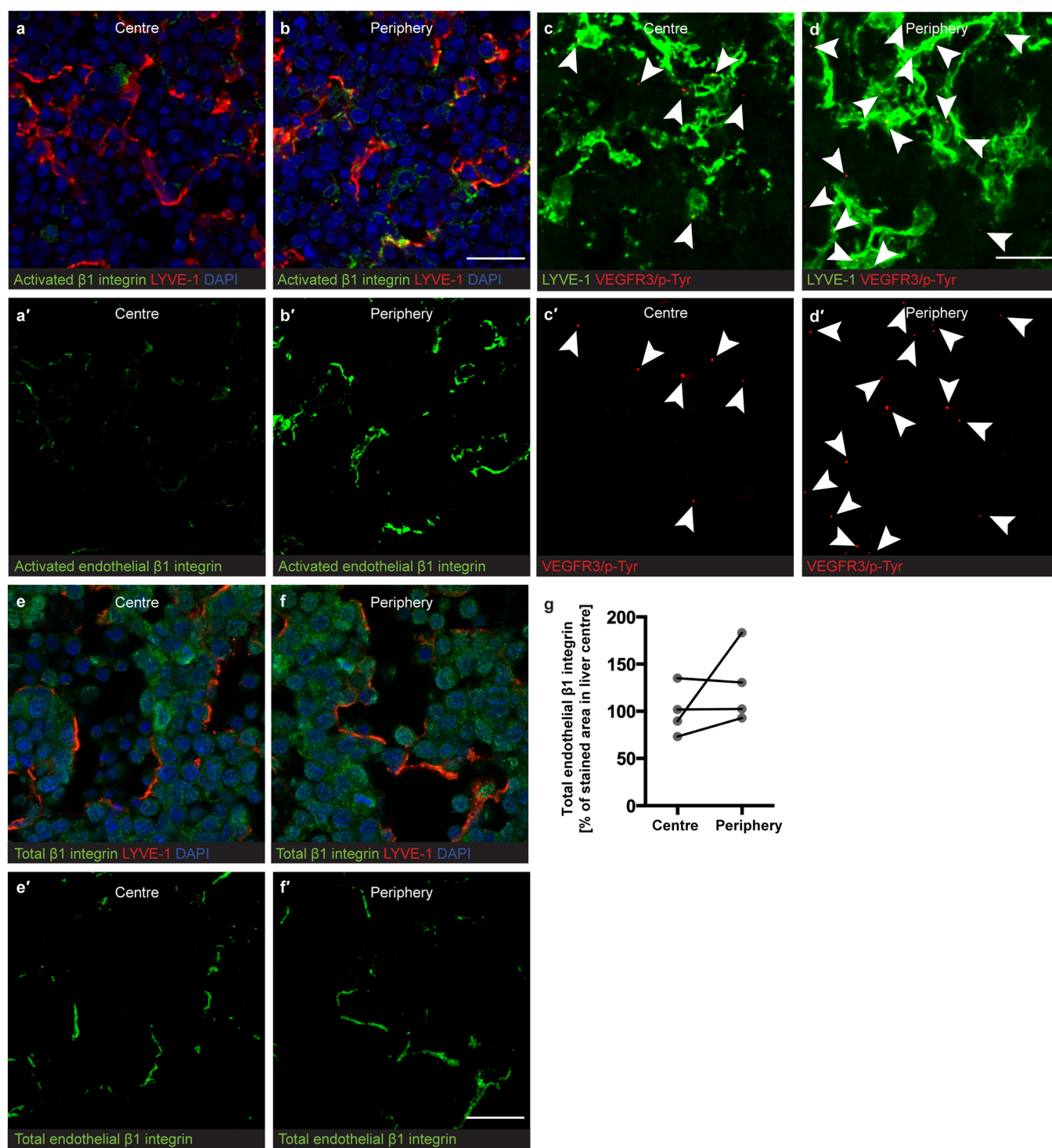
Source Data for all quantifications are provided. Supplementary Video 1 is available at <https://figshare.com/s/7f48df3583bfb7ce0ea1>; Supplementary Video 2 is available at <https://figshare.com/s/361947a6f67b8b925042>; and Supplementary Video 3 is available at <https://figshare.com/s/f7ceb9d5980d10084ad8>. Full scans of western blots are provided in Supplementary Fig. 1.

31. Ventura, A. et al. Restoration of p53 function leads to tumour regression *in vivo*. *Nature* **445**, 661–665 (2007).
32. Haiko, P. et al. Deletion of vascular endothelial growth factor C (VEGF-C) and VEGF-D is not equivalent to VEGF receptor 3 deletion in mouse embryos. *Mol. Cell. Biol.* **28**, 4843–4850 (2008).
33. Zeeb, M. et al. Pharmacological manipulation of blood and lymphatic vascularization in ex vivo-cultured mouse embryos. *Nat. Protocols* **7**, 1970–1982 (2012).
34. Sies, H. The use of perfusion of liver and other organs for the study of microsomal electron-transport and cytochrome P-450 systems. *Methods Enzymol.* **52**, 48–59 (1978).
35. Mitchell, C. & Willenbring, H. A reproducible and well-tolerated method for 2/3 partial hepatectomy in mice. *Nat. Protocols* **3**, 1167–1170 (2008).
36. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C_T method. *Nat. Protocols* **3**, 1101–1108 (2008).
37. Livingstone, R. S. et al. Initial clinical application of modified Dixon with flexible echo times: hepatic and pancreatic fat assessments in comparison with ^1H MRS. *MAGMA* **27**, 397–405 (2014).
38. Häussinger, D. in *Metabolism of Human Diseases* (eds Lammert, E. & Zeeb, M.) 173–180 (Springer, Vienna, 2014).



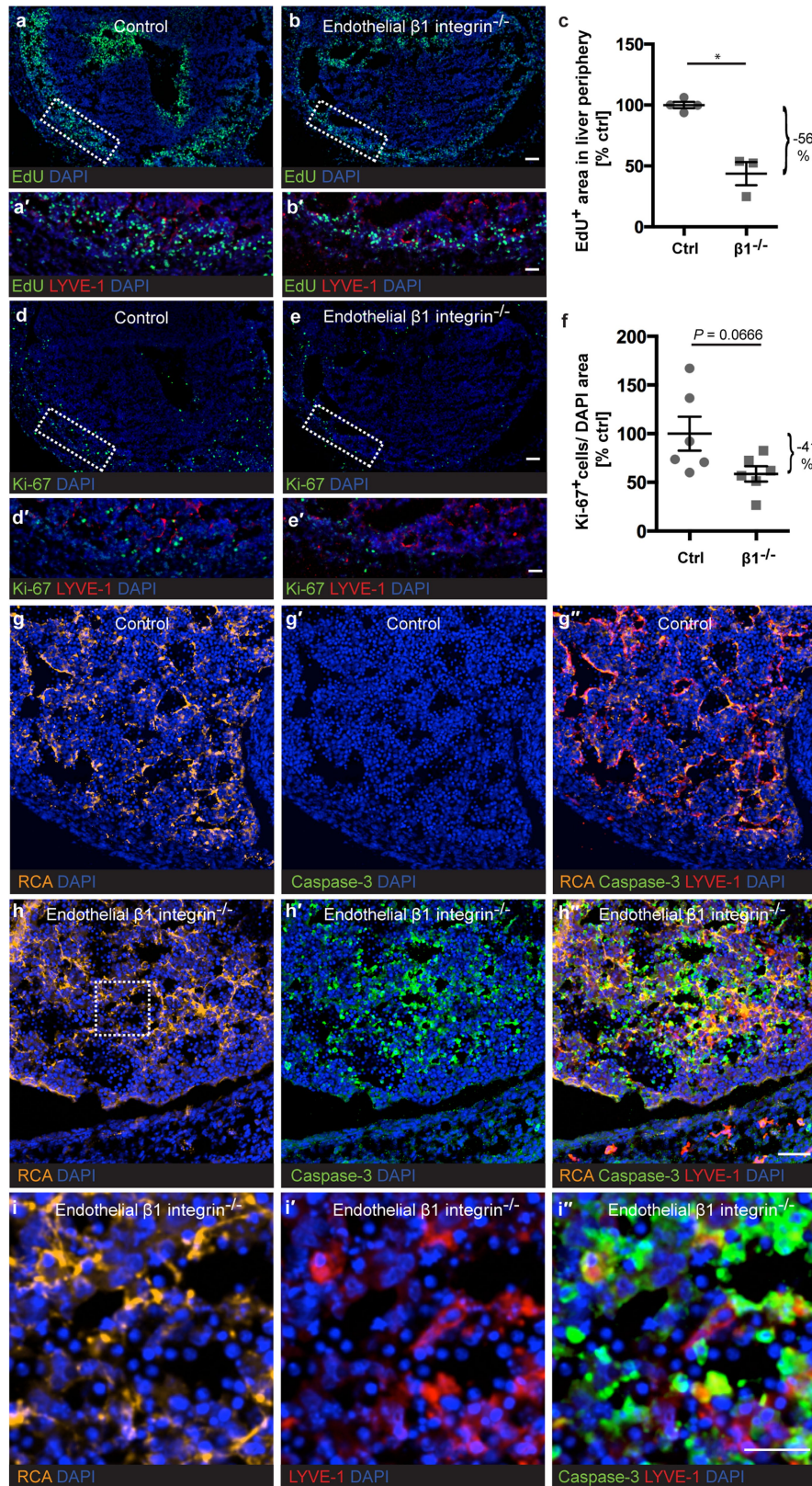
Extended Data Fig. 1 | Hepatic cell proliferation in perfused liver periphery versus liver centre at E11.5. **a, b, c**, Transversal section through a liver after injection of EdU into its lobes and WEC for 3 h. Magnified fields of view in the liver centre (**a'**, **b'**, **c'**) and the liver periphery (**a''**, **b''**, **c''**), as indicated in **c** by white boxes. **d, d'**, **e, e'**, Transversal sections through E11.5 livers showing the liver centre (**d, d'**) and liver

periphery (**e, e'**). **f**, Quantification of the Ki-67⁺ area in the liver centre versus liver periphery ($n = 3$ livers, $*P = 0.0142$ (119.40; 376.70)). Scale bars, 100 μm (**a, b, c**) and 50 μm (all other panels). Two-tailed paired Student's *t*-test, 95% confidence interval (lower confidence limit; upper confidence limit).



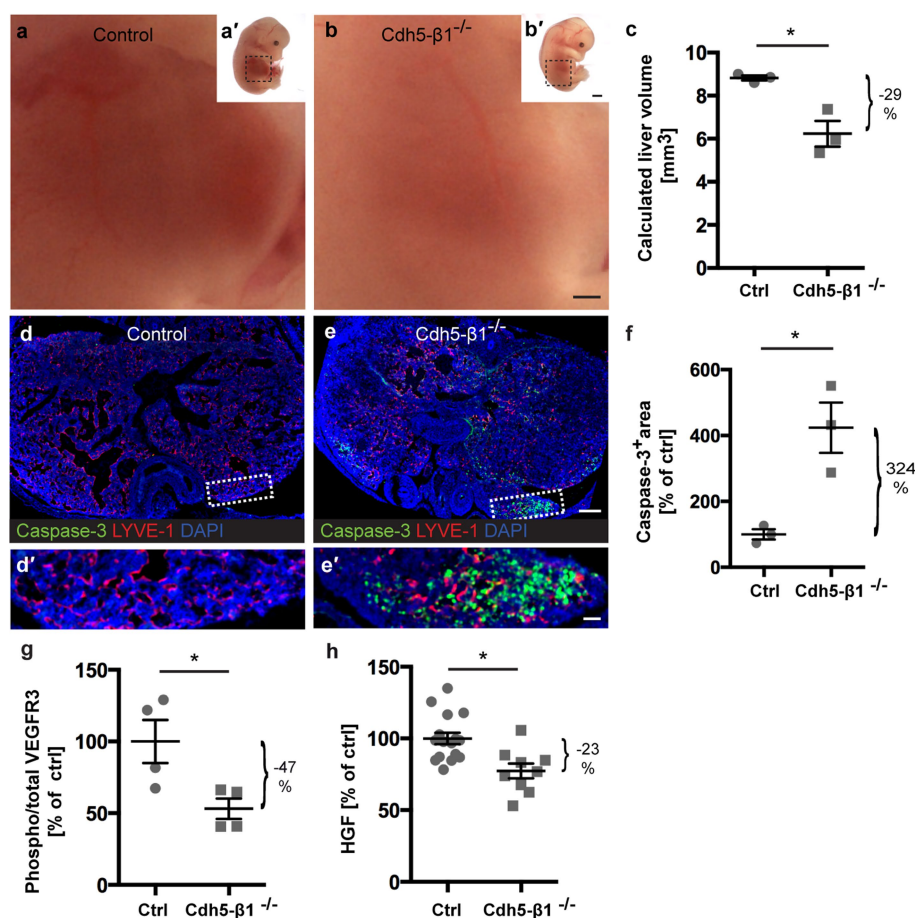
Extended Data Fig. 2 | Activation of $\beta 1$ integrin, VEGFR3 and total $\beta 1$ integrin expression in ECs of E11.5 livers. a, a', b, b', Transversal sections through livers with quantified areas of activated endothelial $\beta 1$ integrin (for quantification, see Fig. 1m). c, c', d, d', Transversal sections with VEGFR3 tyrosine phosphorylation indicated by PLA (red dots and white arrowheads;

for quantification, see Fig. 1n). e, e', f, f', g, Transversal sections with total $\beta 1$ integrin (e, f) and quantified area of total endothelial $\beta 1$ integrin (e', f') in liver centre versus liver periphery (g, $n = 4$ livers, $P = 0.316$ (−44.98; 99.49)). Scale bars, 25 μm . Two-tailed paired Student's t -tests, 95% confidence interval (lower confidence limit; upper confidence limit).



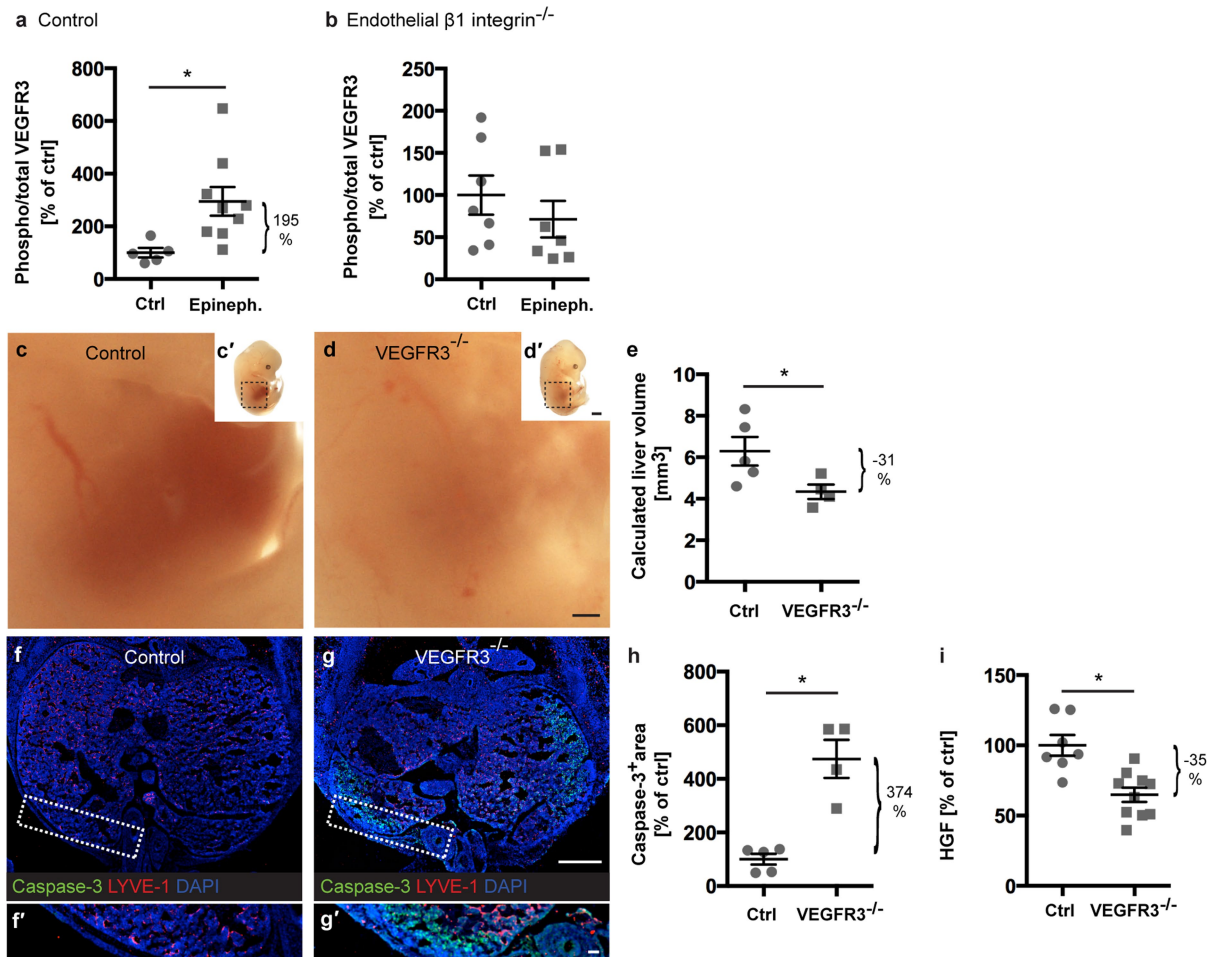
Extended Data Fig. 3 | Endothelial $\beta 1$ integrin is required for hepatic cell proliferation and survival in the liver periphery. **a, a', b, b', d, d', e, e',** Transversal sections through E11.5 livers from EC-specific heterozygous (control) versus homozygous ($\beta 1$ integrin-null) knockouts of $\beta 1$ integrin, including magnifications (**a', b', d', e'**) after 3 h EdU incorporation in WEC. **c,** Quantification of EdU-positive peripheral areas: $n = 4$ control versus $n = 3$ $\beta 1$ integrin-null livers ($*P = 0.0207$ (19.06; 93.54)). **f,** Ki-67-positive cells in embryonic livers: $n = 6$ Cre-control versus

$n = 6$ $\beta 1$ integrin-null livers ($P = 0.0666$ (3.71; -86.50)). **g, g', g'', h, h', h'', i, i', i'',** Transversal sections through control (**g, g', g''**) and $\beta 1$ integrin-null (**h, h', h''**) E13.5 livers with vascular-painted (RCA), perfused hepatic vessels, including magnified images (**i, i', i''**). Scale bars, 100 μm (**a, b, d, e**), 50 μm (**a', b', d', e', g, g', g'', h, h', h''**) and 25 μm (**i, i', i''**). Data are mean \pm s.e.m. Two-tailed unpaired Student's t -tests, 95% confidence interval (lower confidence limit; upper confidence limit).



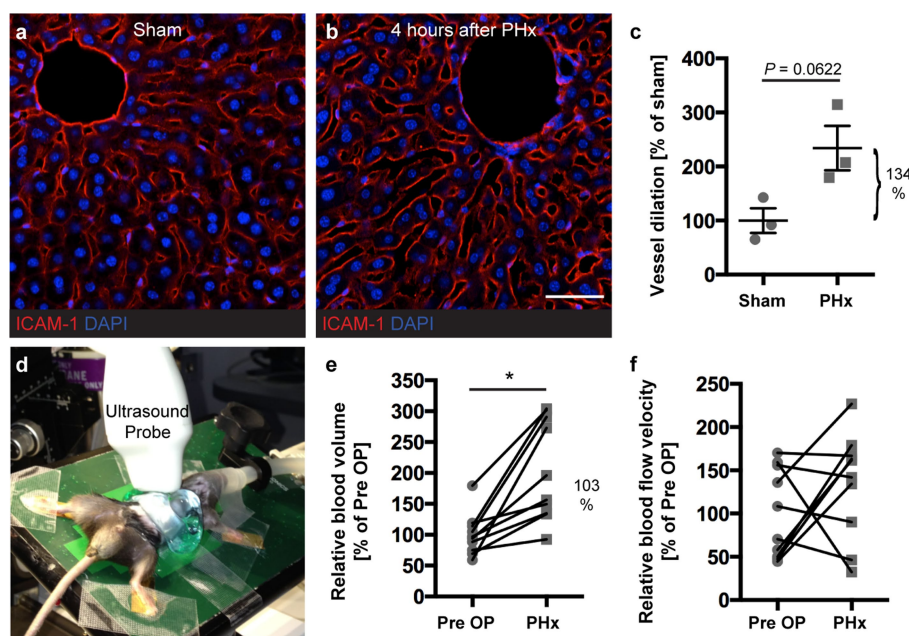
Extended Data Fig. 4 | *Cdh5-cre^{ERT2}*-mediated depletion of $\beta 1$ integrin reduces liver growth, survival, VEGFR3 activation and HGF production. **a, a', b, b', E13.5 mouse embryos (**a', b'**), with their abdomen magnified (**a, b**). **c**, Calculated volumes of livers taken from mouse embryos with *Cdh5-cre^{ERT2}*-mediated depletion of $\beta 1$ integrin (labelled '*Cdh5-β1^{-/-}*') and control littermates ($n = 3$ embryos each, $*P = 0.0446$ (0.15; 5.03)). **d, d', e, e'**, Transversal sections through E13.5 livers with magnifications (**d', e'**). **f**, Quantification of cleaved caspase-3-positive areas ($n = 3$ embryos each, $*P = 0.0462$ (12.62; 634.90)). **g**, VEGFR3**

tyrosine phosphorylation normalized to total VEGFR3 ($n = 4$ control embryos versus $n = 4$ embryos with *Cdh5-cre^{ERT2}*-mediated depletion of $\beta 1$ integrin, $*P = 0.0445$ (1.81; 91.90)). **h**, HGF protein concentrations normalized to total protein ($n = 16$ control embryos versus $n = 9$ embryos with *Cdh5-cre^{ERT2}*-mediated depletion of $\beta 1$ integrin, $*P = 0.0030$ (8.81; 36.46)). Scale bars, 500 μm (**a, b**), 1 mm (**a', b'**), 200 μm (**d, e**) and 50 μm (**d', e'**). Data are mean \pm s.e.m. Two-tailed unpaired Student's *t*-tests, 95% confidence interval (lower confidence limit; upper confidence limit).



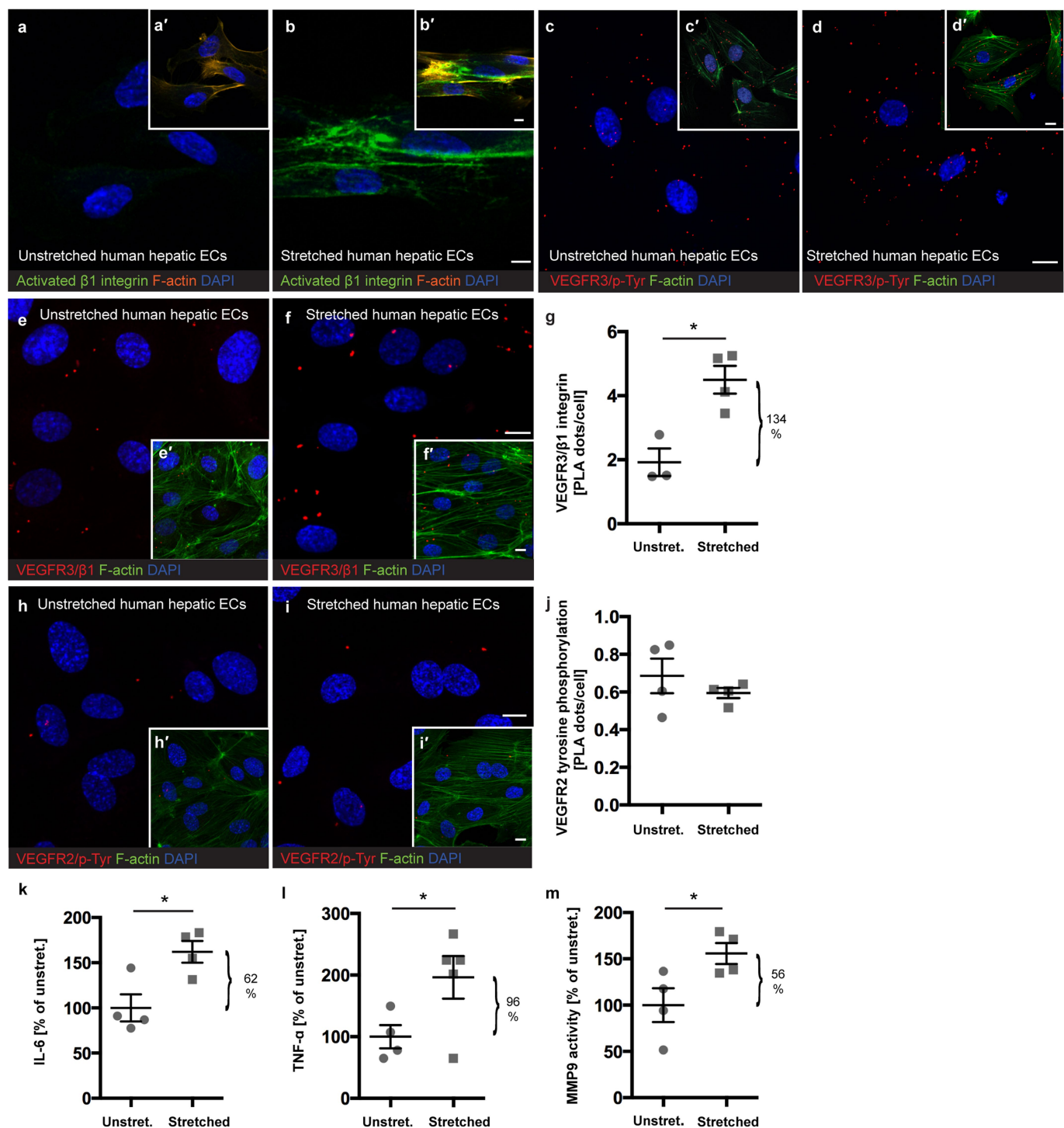
Extended Data Fig. 5 | $\beta 1$ integrin acts upstream of perfusion-dependent activation of VEGFR3, which is required for liver growth and survival. **a, b**, VEGFR3 tyrosine phosphorylation, normalized to total VEGFR3 protein, in liver lysates from E12.5 embryos with heterozygous depletion of endothelial $\beta 1$ integrin (control) (**a**) and E12.5 endothelial $\beta 1$ integrin-null embryos (**b**) in gain-of-perfusion experiments. $n = 5$ control embryos treated with control solution versus $n = 9$ control embryos treated with epinephrine and atropine ($*P = 0.0072$ (66.37;323.00)), and $n = 7$ control embryos treated with control solution versus $n = 7$ epinephrine- and atropine-treated endothelial $\beta 1$ integrin-null embryos, ($P = 0.3853$ (-97.93; 40.64)). **c, c', d, d'**, Abdominal region of an E13.5

control littermate (**c'**) and an E13.5 *Vegfr3*^{-/-} mouse embryo (**d'**). **e**, Calculated liver volumes ($n = 5$ control embryos versus $n = 4$ *Vegfr3*^{-/-} embryos, $*P = 0.0462$ (0.05; 3.86)). **f, f', g, g'**, Transversal sections through E13.5 livers with magnifications (**f', g'**). **h**, Quantification of cleaved caspase-3-positive areas: $n = 5$ control livers versus $n = 4$ *Vegfr3*^{-/-} livers ($*P = 0.0103$ (156.50; 591.60)). **i**, HGF protein concentrations in lysates from embryonic livers normalized to total protein: $n = 7$ control embryos versus $n = 10$ *Vegfr3*^{-/-} embryos ($*P = 0.0022$ (15.55; 54.81)). Scale bars, 500 μm (**c, d, f, g**), 1 mm (**c', d'**) and 50 μm (**f', g'**). Data are mean \pm s.e.m. Two-tailed unpaired Student's *t*-tests, 95% confidence interval (lower confidence limit; upper confidence limit).



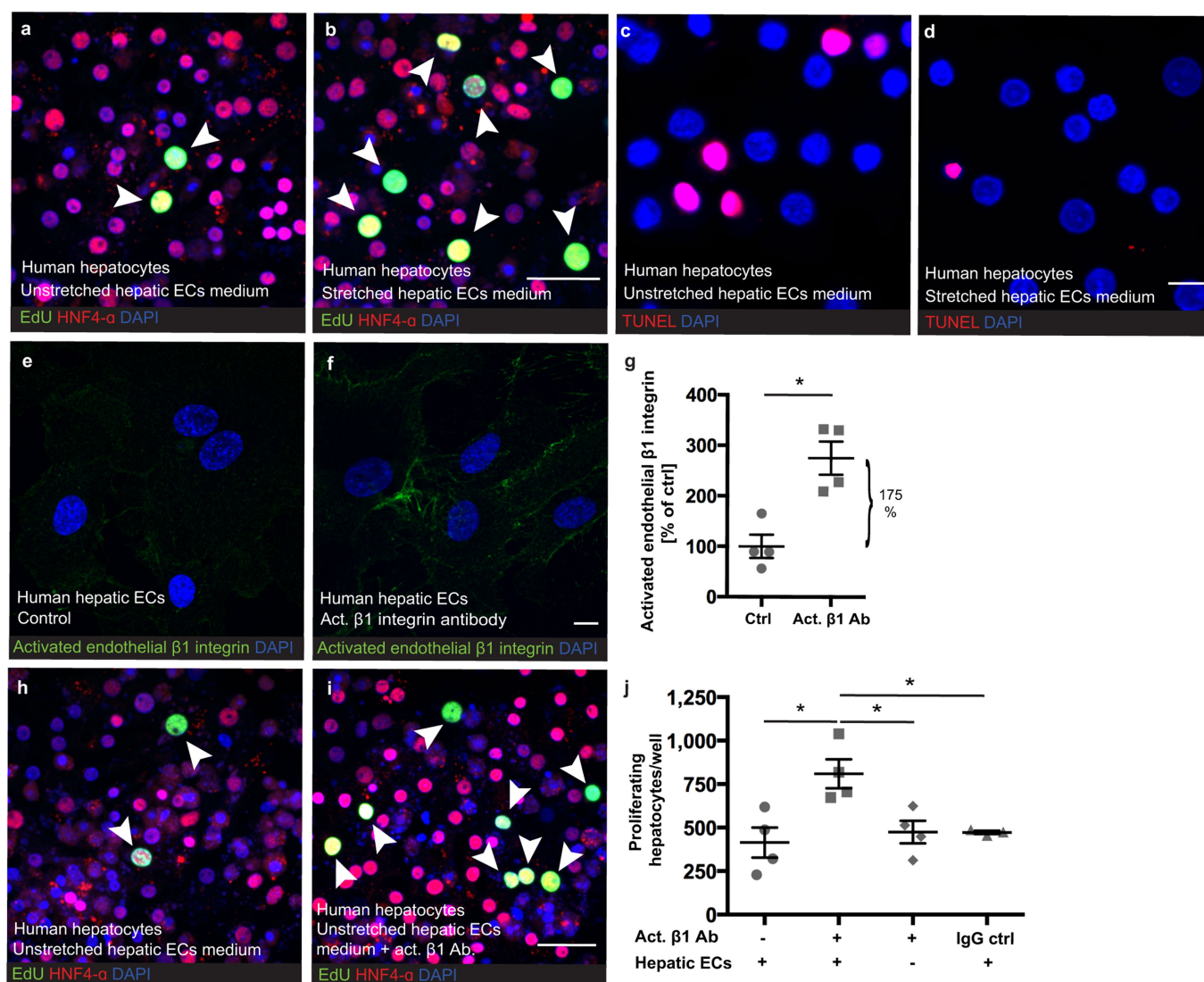
Extended Data Fig. 6 | Vessel dilation, blood volume and blood flow velocity in liver sinusoids after a two-thirds partial hepatectomy. **a–c**, Sections through livers isolated from adult mice after sham operation (**a**) versus partial hepatectomy (PHx) (**b**), and quantification (**c**) of vessel dilation by measurement of vascular lumen areas ($n = 3$ livers each, $P = 0.0622$ (-12.43 ; 280.20)). **d**, Contrast-enhanced ultrasound measurements on the liver of a mouse. **e**, **f**, Relative blood volume (calculated from peak enhancement of contrast agent, $n = 10$ mice,

$*P = 0.0014$ (51.98 ; 154.40)) (**e**) and relative blood flow velocity (calculated from time to peak of contrast agent enhancement, $n = 10$ mice, $P = 0.2264$ (-25.41 ; 93.88)) (**f**) in the right liver lobe of adult mice normalized to the liver area before ('Pre OP') and after partial hepatectomy. Data in **c** are mean \pm s.e.m. Scale bar, $20\ \mu\text{m}$. Student's t -tests were two-tailed unpaired (**c**) or two-tailed paired (**e**, **f**), 95% confidence interval (lower confidence limit; upper confidence limit).



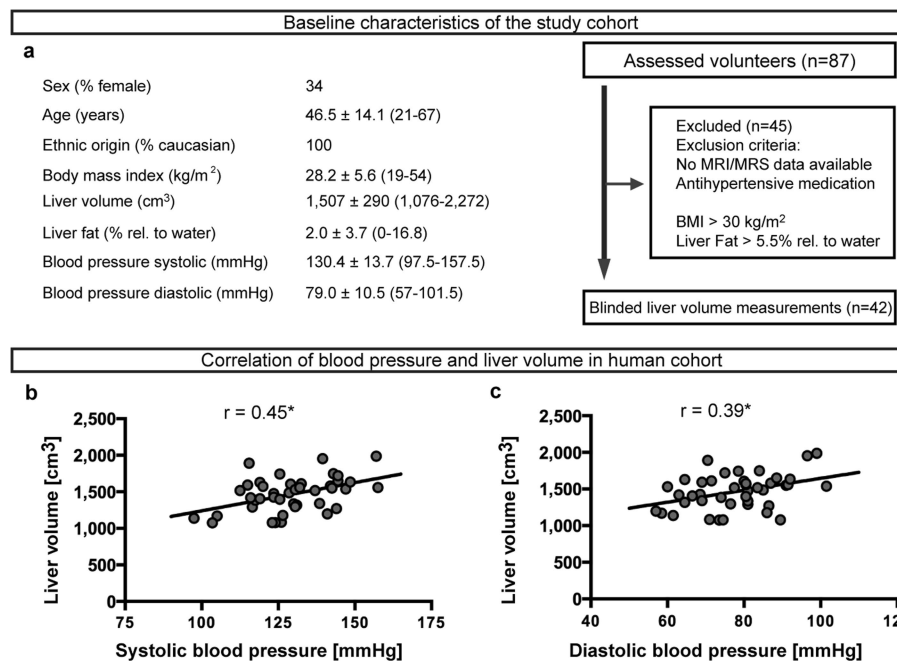
Extended Data Fig. 7 | Mechanically induced release of angiocrine signals. **a, a', b, b', c, c', d, d', e, e', f, f', h, h', i, i',** Unstretched versus mechanically stretched human hepatic ECs, stained for F-actin (**a', b', c', d', e', f', h', i',**), activated $\beta 1$ integrin (**a, a', b, b'**) and shown as PLA dots: VEGFR3 tyrosine phosphorylation (**c, c', d, d'**), co-localization of $\beta 1$ integrin and VEGFR3 (**e, e', f, f'**) and VEGFR2 tyrosine phosphorylation (**h, h', i, i'**). **g, j,** Quantification of the interaction between VEGFR3 and $\beta 1$ integrin ($n = 3$ unstretched chambers versus $n = 4$ stretched chambers, $*P = 0.0091$ (0.99; 4.16)), and VEGFR2 tyrosine phosphorylation

($n = 4$ stretch chambers each, $P = 0.4000$ (−0.37; 0.19)). **k,** IL-6 protein concentration ($n = 4$ stretch chambers each, $*P = 0.0191$ (14.51; 109.70)). **l,** TNF protein concentration ($n = 4$ unstretched chambers versus $n = 5$ stretched chambers, $*P = 0.0493$ (0.43; 192.40)). **m,** MMP9 activity ($n = 4$ stretch chambers each, $*P = 0.0488$ (0.44; 111.20)), in the supernatant of unstretched versus stretched hepatic ECs. Scale bars, 10 μm . Data are mean \pm s.e.m. Two-tailed unpaired Student's t -tests, 95% confidence interval (lower confidence limit; upper confidence limit).



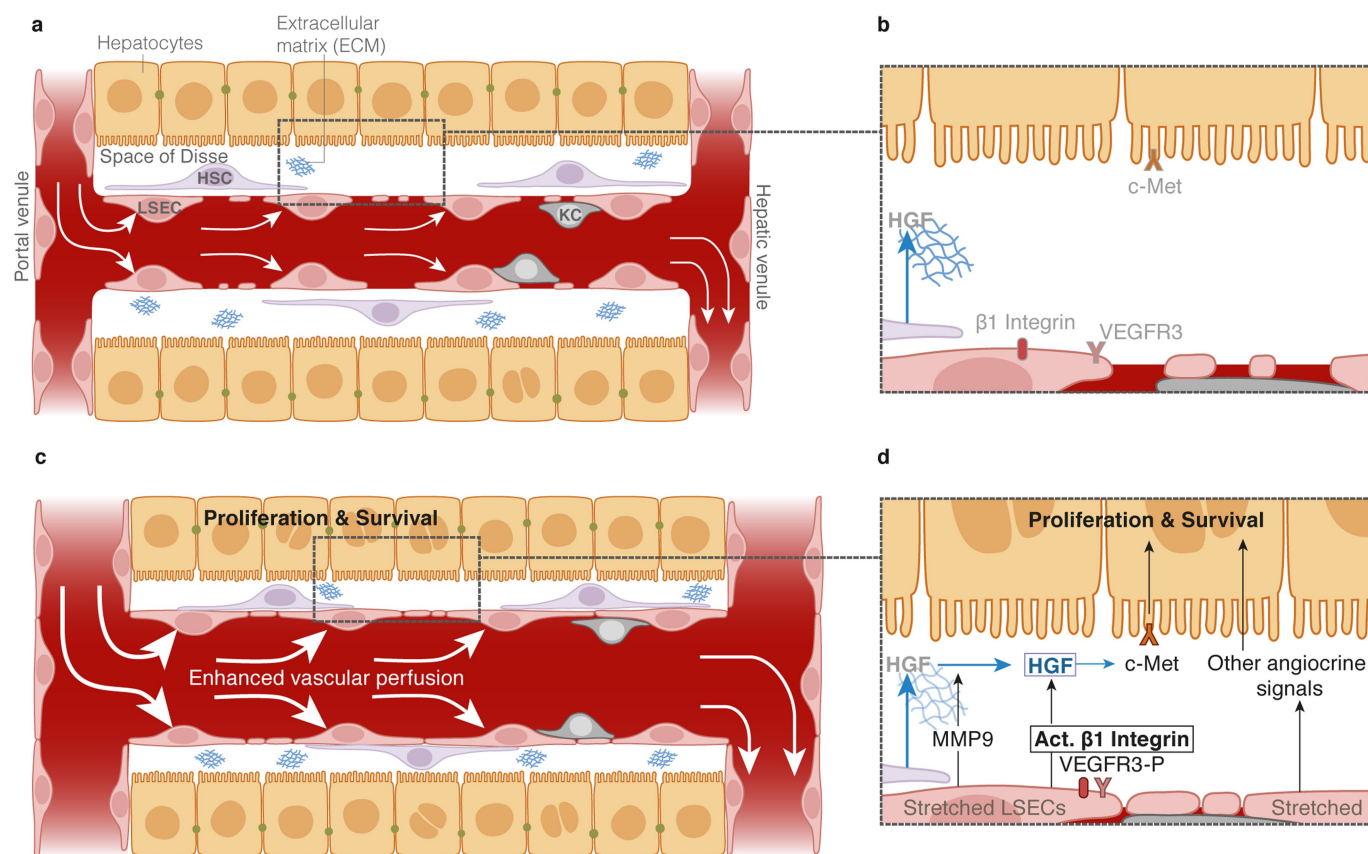
Extended Data Fig. 8 | Mechanically induced angiocrine signals promote proliferation and survival of human hepatocytes. **a–d**, Primary human hepatocytes treated for 6 h with supernatant from unstretched (**a**, **c**) versus stretched (**b**, **d**) human hepatic ECs. **e–g**, Unstretched ECs without (**e**) or with (**f**) activating $\beta 1$ integrin antibody, including quantification (**g**) of $\beta 1$ integrin activation ($n = 4$ wells each, $*P = 0.0062$ (73.97; 276.10)). **h**, **i**, Primary human hepatocytes treated for 6 h with supernatant from unstretched hepatic ECs treated without (**h**) or with (**i**) an activating $\beta 1$ integrin antibody. **j**, Quantification of hepatocyte proliferation after

incubation with supernatant from ECs without or with activating $\beta 1$ integrin antibody ($n = 4$ wells each, $*P = 0.0110$ (91.40; 698.60)), with activating $\beta 1$ integrin antibody alone without previous incubation with ECs ($n = 4$ wells, $*P = 0.0299$ (31.15; 638.40)) and with conditioned medium of ECs treated with an IgG₁ isotype control ($n = 3$ wells, $*P = 0.0436$ (8.99; 664.80)). Scale bars, 50 μm (**a**, **b**, **h**, **i**) and 10 μm (**c**, **d**, **e**, **f**). Data are mean \pm s.e.m. Two-tailed unpaired Student's t -tests (**g**) and one-way ANOVA followed by Tukey's test (**j**), 95% confidence interval (lower confidence limit; upper confidence limit).



Extended Data Fig. 9 | Correlation of blood pressure and liver volume in metabolically healthy human individuals. **a**, Baseline characteristics (mean ± standard deviation, minimum and maximum values) of the study cohort, which was comprised of individuals recruited from glucose-tolerant humans who served as controls in the prospective observational German Diabetes Study²⁹. **b**, **c**, Correlation of blood pressure with liver

volume as assessed by MRI taken between 07:00 and 10:30. Correlation of systolic blood pressure with liver volume ($n = 42$; $r = 0.45$; $*P = 0.0022$ (0.17; 0.66)), and correlation of diastolic blood pressure with liver volume ($n = 42$; $r = 0.39$; $*P = 0.0104$ (0.09; 0.62)). Reported correlations are Pearson correlation coefficients with 95% confidence intervals (lower confidence limit; upper confidence limit).



Extended Data Fig. 10 | Model of mechanotransduced angiocrine signals in the liver. Simplified drawings of liver sinusoids with liver sinusoidal ECs (LSECs), hepatic stellate cells (HSCs), extracellular matrix (ECM), Kupffer cells (KCs) and hepatocytes (with schematic elements taken from a previous publication³⁸). **a**, Liver sinusoid under normal blood flow. **b**, Magnification of space of Disse. **c**, Liver sinusoid with enhanced blood perfusion. **d**, Magnification of space of Disse. When the vascular

lumen widens owing to enhanced blood perfusion, circumferential stretching of liver sinusoidal ECs activates endothelial $\beta 1$ integrin and its interaction with VEGFR3. The hepatic ECs (in concert with other cells such as HSCs) subsequently release angiocrine signals—such as HGF, IL-6 and TNF—and activate MMP9, and thus enhance proliferation and survival of the adjacent hepatocytes.

Discovery of a periosteal stem cell mediating intramembranous bone formation

Shawon Debnath¹, Alisha R. Yallowitz¹, Jason McCormick², Sarfaraz Lalani¹, Tuo Zhang³, Ren Xu¹, Na Li¹, Yifang Liu⁴, Yeon Suk Yang⁵, Mark Eiseman¹, Jae-Hyuck Shim⁵, Meera Hameed⁶, John H. Healey⁷, Mathias P. Bostrom^{8,9}, Dan Avi Landau^{10,11} & Matthew B. Greenblatt^{1*}

Bone consists of separate inner endosteal and outer periosteal compartments, each with distinct contributions to bone physiology and each maintaining separate pools of cells owing to physical separation by the bone cortex. The skeletal stem cell that gives rise to endosteal osteoblasts has been extensively studied; however, the identity of periosteal stem cells remains unclear^{1–5}. Here we identify a periosteal stem cell (PSC) that is present in the long bones and calvarium of mice, displays clonal multipotency and self-renewal, and sits at the apex of a differentiation hierarchy. Single-cell and bulk transcriptional profiling show that PSCs display transcriptional signatures that are distinct from those of other skeletal stem cells and mature mesenchymal cells. Whereas other skeletal stem cells form bone via an initial cartilage template using the endochondral pathway⁴, PSCs form bone via a direct intramembranous route, providing a cellular basis for the divergence between intramembranous versus endochondral developmental pathways. However, there is plasticity in this division, as PSCs acquire endochondral bone formation capacity in response to injury. Genetic blockade of the ability of PSCs to give rise to bone-forming osteoblasts results in selective impairments in cortical bone architecture and defects in fracture healing. A cell analogous to mouse PSCs is present in the human periosteum, raising the possibility that PSCs are attractive targets for drug and cellular therapy for skeletal disorders. The identification of PSCs provides evidence that bone contains multiple pools of stem cells, each with distinct physiologic functions.

A major limitation to identifying periosteal stem cells has been the lack of genetic markers that discriminate between periosteal and endosteal mesenchyme. While studying the specificity of skeletal targeting *cre* strains, we observed that cathepsin K-Cre (*Ctsk^{cre}*) labels the periosteal mesenchyme. In *Ctsk^{cre};Rosa26^{mT/mG}* reporter mice⁶, in which *Ctsk^{cre}* cells and their progeny (hereafter CTSK–mGFP cells) express membrane-bound GFP (mGFP), labelling of the periosteal mesenchyme was observed as early as embryonic day 14.5 (E14.5) (Extended Data Fig. 1a–e). At postnatal day 10, CTSK–mGFP cells were observed in the periosteal mesenchyme and the endosteal marrow compartment, although nearly all of the endosteal cells were morphologically consistent with osteoclasts (Fig. 1a, Extended Data Fig. 1f). A negligible number of osteocytes were CTSK–mGFP⁺ (Extended Data Fig. 1g). Flow cytometry of endosteal cells and co-staining for tartrate-resistant acid phosphatase (TRAP) confirmed that endosteal CTSK–mGFP cells were osteoclasts (Fig. 1b, c). Conversely, the majority of CTSK–mGFP cells in the periosteum were CD45[–]TER119[–]CD31[–] (hereafter Lin[–]) mesenchymal cells (Fig. 1c). Periosteal CTSK–mGFP cells include periosteal osteoblasts, as shown by expression of type I collagen, Runx2, alkaline phosphatase (ALPL) and osteocalcin (Fig. 1d, Extended Data

Fig. 1j). Therefore, within the mesenchymal compartment, *Ctsk^{cre}* selectively labels the periosteum⁷ (Fig. 1c).

This labelling suggested that a periosteal stem cell exists within the mesenchymal CTSK–mGFP⁺ population (Fig. 1a–c). To test this, we fractionated CTSK–mGFP mesenchymal cells using multi-colour flow cytometry (Fig. 1f–j, Extended Data Fig. 1i). We observed three populations among CTSK–mGFP cells lacking THY1.2 and 6C3⁸, all of which were CD49^{low}CD51^{low}; CD200⁺CD105[–] periosteal mesenchymal stem cells (PSCs), CD200[–]CD105[–] periosteal progenitor 1 (PP1) cells, and CD105⁺CD200^{variable} periosteal progenitor 2 (PP2) cells (Fig. 1f, g, Extended Data Fig. 2a–c). Immunostaining confirmed the presence of CD200⁺ CTSK–mGFP cells in the periosteum and also identified subsets of CTSK–mGFP cells expressing gremlin¹⁵ and nestin¹ (Fig. 1e, Extended Data Fig. 1b, e, k). Consistent with the restriction of haematopoiesis to the endosteal compartment, CTSK–mGFP⁺ status, and expression of LEPR, CD146 and CD140α—markers that are present in mesenchymal cells and have the ability to support haematopoiesis^{2,3,9,10}—are mutually exclusive (Fig. 1h–l, Extended Data Fig. 2e–g). During embryonic development, periosteal CD200⁺ CTSK–mGFP cells consistent with PSCs were first observed at E14.5, concurrent with the onset of skeletal mineralization, and were distinct from CD200⁺ cells present within the chondroepiphysis (Extended Data Fig. 1a–e).

PSCs display the stem cell properties of clonal multipotency, self-renewal and the ability to give rise to the entire range of CTSK–mGFP cells. Serial mesosphere formation is an in vitro proxy for self-renewal¹, and only PSCs possessed the capacity to form tertiary mesospheres, retaining CD200 through this process (Fig. 2a–c). To determine which periosteal population sits at the apex of the CTSK–mGFP differentiation hierarchy, cell populations isolated by fluorescence-activated cell sorting (FACS) were cultured for 15 days, and subsequently reanalysed by flow cytometry; these PSCs differentiated into PP1 and PP2 cells in addition to THY1.2⁺ and 6C3⁺ cells (Fig. 2d). By contrast, PP1s or PP2s did not produce PSCs in culture (Extended Data Fig. 2i). Additionally, PSCs demonstrated in vitro clonal multipotency for differentiation into mature osteoblasts and adipocytes (Fig. 2e). Similarly, a clonogenic periosteal population can be identified by pulse labelling in vivo (Extended Data Fig. 2h). PSCs also possessed the capacity to differentiate into chondrocytes (Fig. 2e). In summary, PSCs are the most stem-like of the CTSK–mGFP populations in vitro.

In contrast to other skeletal mesenchymal stem cells (MSCs) that mediate endochondral ossification, PSCs form bone in vivo via an intramembranous pathway. We transplanted PSCs and non-CTSK (Lin[–]) MSCs under the kidney capsule of wild-type secondary recipient mice (Fig. 2f, g). Both non-CTSK MSCs and PSCs mediated de novo generation of bone organoids (Fig. 2h). Consistent with the physiologic restriction of marrow recruitment to the endosteal compartment,

¹Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. ²Flow Cytometry Core Facility, Weill Cornell Medicine, New York, NY, USA. ³Genomics Resources Core Facility, Weill Cornell Medicine, New York, NY, USA. ⁴Pathology and Laboratory Medicine Core Facility, Weill Cornell Medicine, New York, NY, USA. ⁵Department of Medicine, University of Massachusetts Medical School, North Worcester, MA, USA. ⁶Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁷Orthopaedic Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁸Research Division, Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, NY, USA. ⁹Division of Adult Reconstruction and Joint Replacement, Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, NY, USA. ¹⁰Cancer Genomics and Evolutionary Dynamics, Weill Cornell Medicine, New York, NY, USA. ¹¹New York Genome Center, New York, NY, USA. *e-mail: mag3003@med.cornell.edu

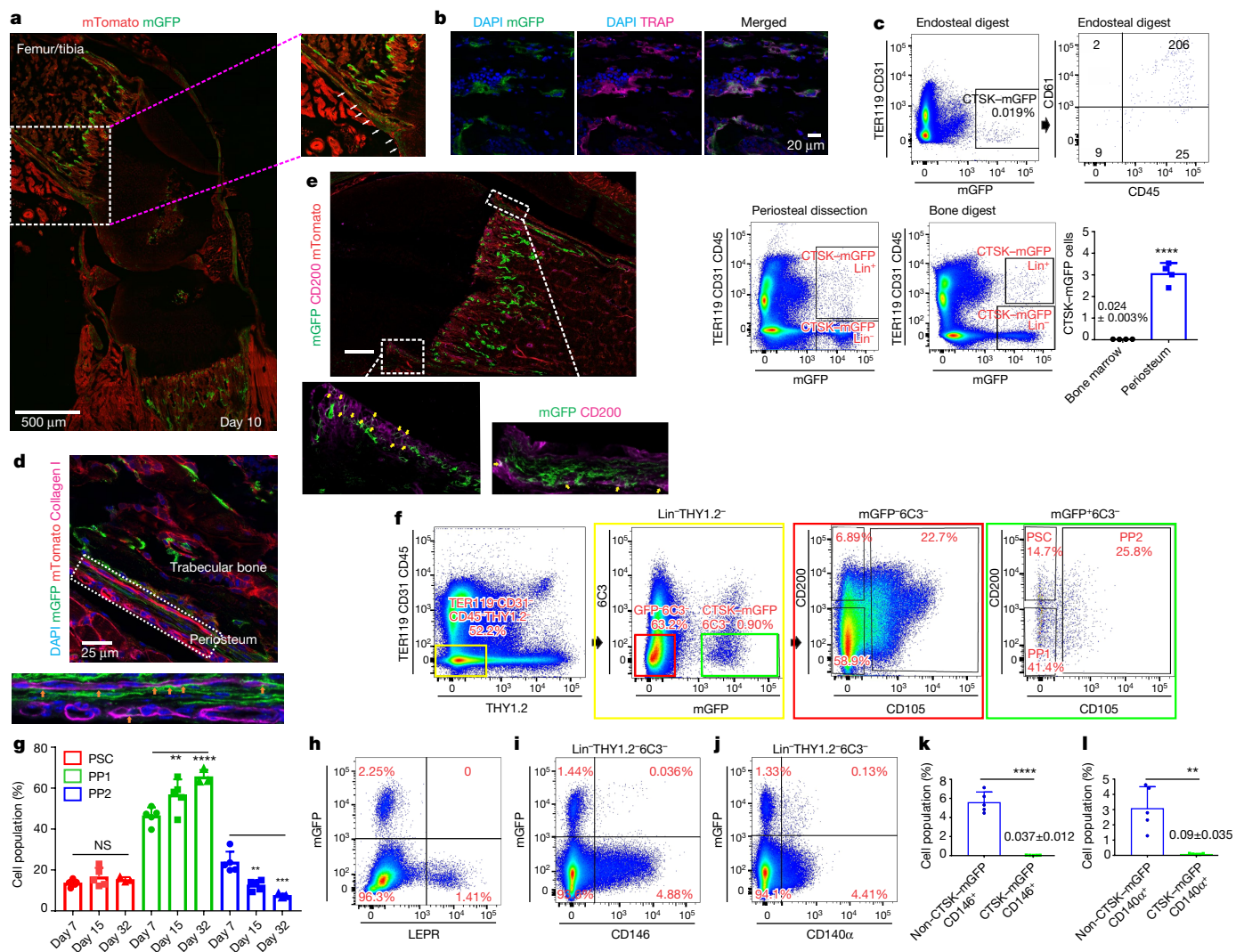


Fig. 1 | Cathepsin K-Cre labels periosteal mesenchymal cells. **a**, Left, mGFP (green) signal in femur/tibia of *Ctsk^{Cre};Rosa26^{mT/mG}* mice at postnatal day 10 (P10). Scale bar, 500 μ m. Right, enlarged view of dotted white box in main panel. Arrows indicate the CTSK-mGFP signal in the periosteum. **b**, Endosteal CTSK-mGFP cells express the osteoclast marker TRAP (magenta). Scale bar, 20 μ m. **c**, Representative FACS analysis of distribution of CTSK-mGFP cells in the endosteal digest, periosteum and total bone digests. **** $P = 1.69 \times 10^{-5}$, two-tailed Student's *t*-test; data are mean \pm s.d.; $n = 4$ independent experiments. **d**, **e**, A subset of CTSK-mGFP (green) periosteal cells in the femur expresses type 1 collagen (d; magenta; orange arrow; bottom panel shows enlarged view of outlined area), and a subset expresses CD200 (e; magenta; yellow arrows; panels on the right show enlarged views of outlined areas). Scale bar, 25 μ m (d, top), 200 μ m (e, top). An enlarged view of **e** is presented in Extended Data Fig. 1h. **f**, Flow cytometry of cells from long bone digests from P7

non-CTSK MSCs underwent endochondral ossification with cartilage differentiation and recruitment of haematopoietic elements, whereas PSCs mediated intramembranous bone formation without haematopoietic recruitment or cartilage formation¹¹ (Fig. 2g–i, Extended Data Fig. 2j–l). Otherwise, non-CTSK MSCs displayed similar performance to PSCs in bone organoid, clonal multipotency and mesosphere assays (Extended Data Fig. 3a–h). No interconversion between non-CTSK MSCs and PSCs was observed after transplantation (Extended Data Fig. 3i–k). Therefore, bone contains discrete stem cell populations with differing functional specializations^{1,2,4,5,12}, including a PSC population specialized for periosteal physiology.

Additionally, PSCs sit at the apex of a differentiation hierarchy in vivo and are capable of self-renewal during serial transplantation

mice to identify periosteal stem cells (PSC) and progenitor cells (PP1, PP2). Colour-coded boxes (yellow, red, green) indicate parent/daughter gates. **g**, Percentage of PSC, PP1 and PP2 populations over time in long bone digests. PP1: ** $P = 0.0063$ (day 15), **** $P = 0.0001$ (day 32). PP2: ** $P = 0.0022$ (day 15), **** $P = 0.0001$ (day 32). One-way ANOVA, Sidak's multiple comparison test. Data are mean \pm s.d.; $n = 5$ (days 7 and 15), $n = 3$ (day 32); representative of 3 independent experiments. **h–j**, Flow cytometry for LEPR (**h**), CD146 (**i**) and CD140 α (**j**) versus CTSK-mGFP in long bones. **k**, **l**, There are significantly fewer CD146⁺ (**k**; **** $P = 0.0000029$) and CD140 α ⁺ (**l**; ** $P = 0.0014$) cells among CTSK-mGFP cells than among non-CTSK-mGFP cells. Two-tailed Student's *t*-test. Data are mean \pm s.d.; 3 independent experiments. Images are representative of 5 (**a**, **b**) or 3 (**d**, **e**) independent experiments. Plots are representative of $n = 20$ (**f**) or $n = 10$ (**h–j**) independent experiments.

assays^{1,9}. CTSK-mGFP PSCs were transplanted into the mammary fat pad of female *Rosa26^{mT/mG}* mice and analysed over two successive rounds of transplantation (Fig. 2j). After the first round of transplantation, PSCs both self-renewed and gave rise to the entire spectrum of CTSK-mGFP cells observed in native periosteum (Fig. 2k, top panels). Similar results were observed in a kidney capsule system (Extended Data Fig. 4a). PSCs from secondary hosts were re-isolated and transplanted into tertiary hosts where they again displayed intact self-renewal and differentiation capacity (Fig. 2k, bottom panels). By contrast, CTSK-mGFP⁺ PP1 or PP2 cells did not revert to PSCs after transplantation and were unable to maintain THY1.2⁺, 6C3⁺ progenitor cells (Extended Data Figs. 3l–n, 4b, c). Therefore, PSCs retain both self-renewal and differentiation capacity through successive rounds of transplantation,

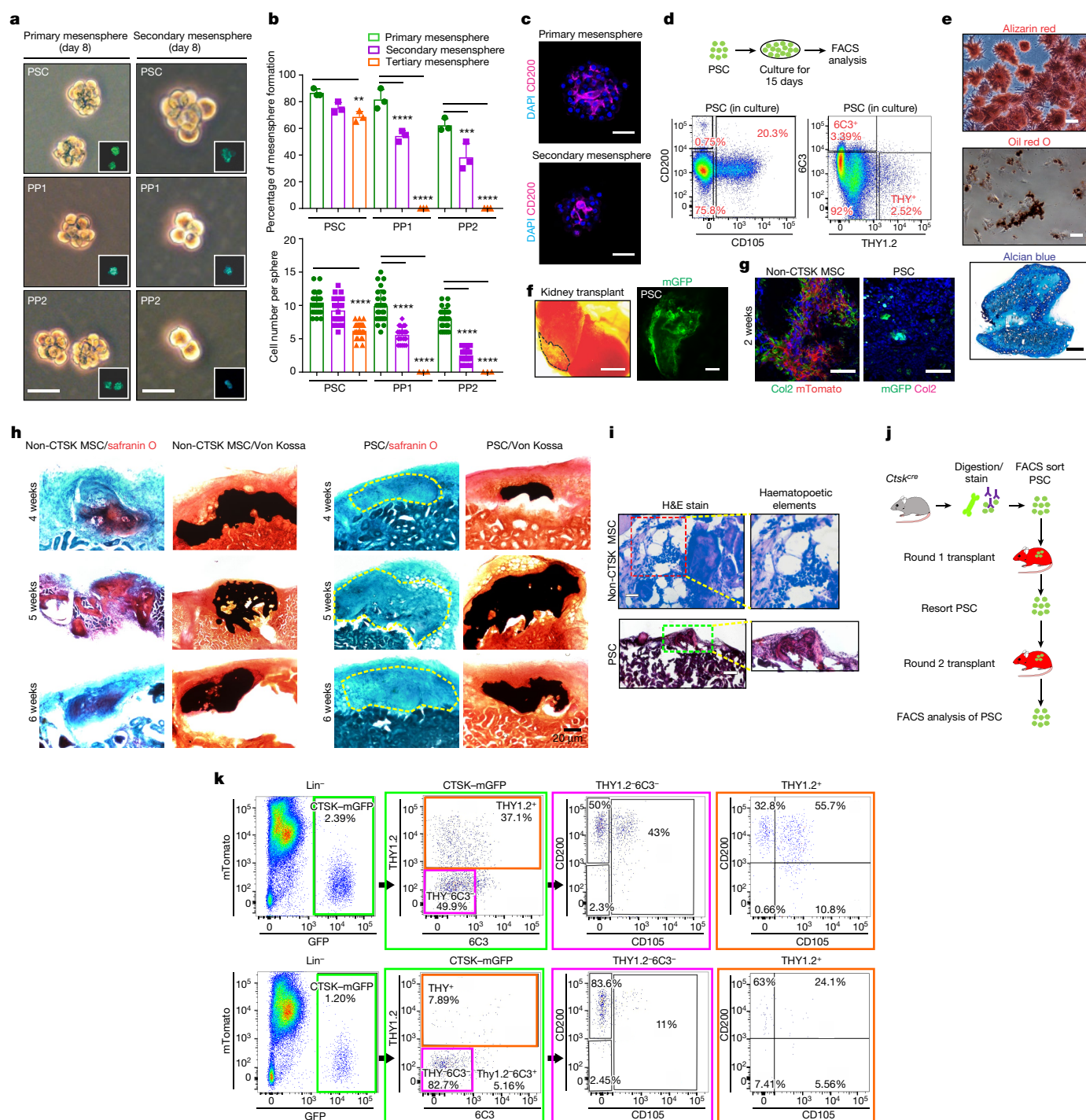


Fig. 2 | Functional characterization of periosteal stem cells. **a**, Bright-field images, primary (left; scale bar, 20 μ m) and secondary (right; scale bar, 10 μ m) mesospheres derived from CTSK-mGFP PSCs (top), PP1s (middle) and PP2s (bottom). GFP (green) in inset. **b**, Percentage of cells able to form spheres (top) and cell number per sphere (bottom) for PSCs, PP1s and PP2s. $**P = 0.0036$, $***P = 0.0002$, $****P = 0.0001$. Dunnett's multiple comparison test. Data are mean \pm s.d., $n = 3$ independent experiments. **c**, Immunostaining for CD200 (magenta) in primary (top) and secondary (bottom) PSC-derived mesospheres. Scale bars, 100 μ m. **d**, Sorted PSCs were cultured for 15 days and analysed by FACS. **e**, PSC colonies, derived from single cells, were split for differentiation into osteoblasts (alizarin red staining, top; scale bar, 50 μ m), adipocytes (oil red O staining, middle; scale bar, 50 μ m). Chondrocyte differentiation potential of these colonies was assayed separately (alcian blue staining, bottom; scale bar, 200 μ m). **f-i**, CTSK⁻ MSCs and CTSK-mGFP PSCs were transplanted into kidney capsule of secondary recipients (f, left; dotted black outline; scale bar, 1 mm). Donor origin of PSCs was

confirmed by GFP retention (f, right; scale bar, 50 μ m). **g**, Immunostaining for type 2 collagen (green in left panel; magenta in right panel) in organoids derived from two-week non-CTSK MSCs (red, left; scale bar, 50 μ m) or PSCs (green, right; scale bar, 100 μ m). **h**, Safranin O staining (red) for cartilage and von Kossa staining (black) for mineralized bone in organoids derived from non-CTSK MSCs or PSCs, 4 (top), 5 (middle) or 6 weeks (bottom) after transplantation. Scale bar, 20 μ m. Yellow coloured outlines correspond to the transplanted tissue. **i**, Haematoxylin and eosin (H&E) staining of organoids derived from non-CTSK MSCs (top) or PSCs (bottom). Scale bars, 10 μ m (top), 20 μ m (bottom). Right, enlarged view of the outlined region. **j**, Schematic of serial transplantation of PSCs into mouse mammary fat pad. **k**, FACS plots of PSC-derived cells after the first (top panels) and second (bottom panels) round of transplantation. Lin⁻ cells are Ter119⁻CD45⁻CD31⁻. Colour-coded boxes (green, magenta, orange) indicate parent/daughter gates. Images are representative of 3 (a, c, e) or 8 (f-i) independent experiments. Plots in d and k are representative of 3 independent experiments.

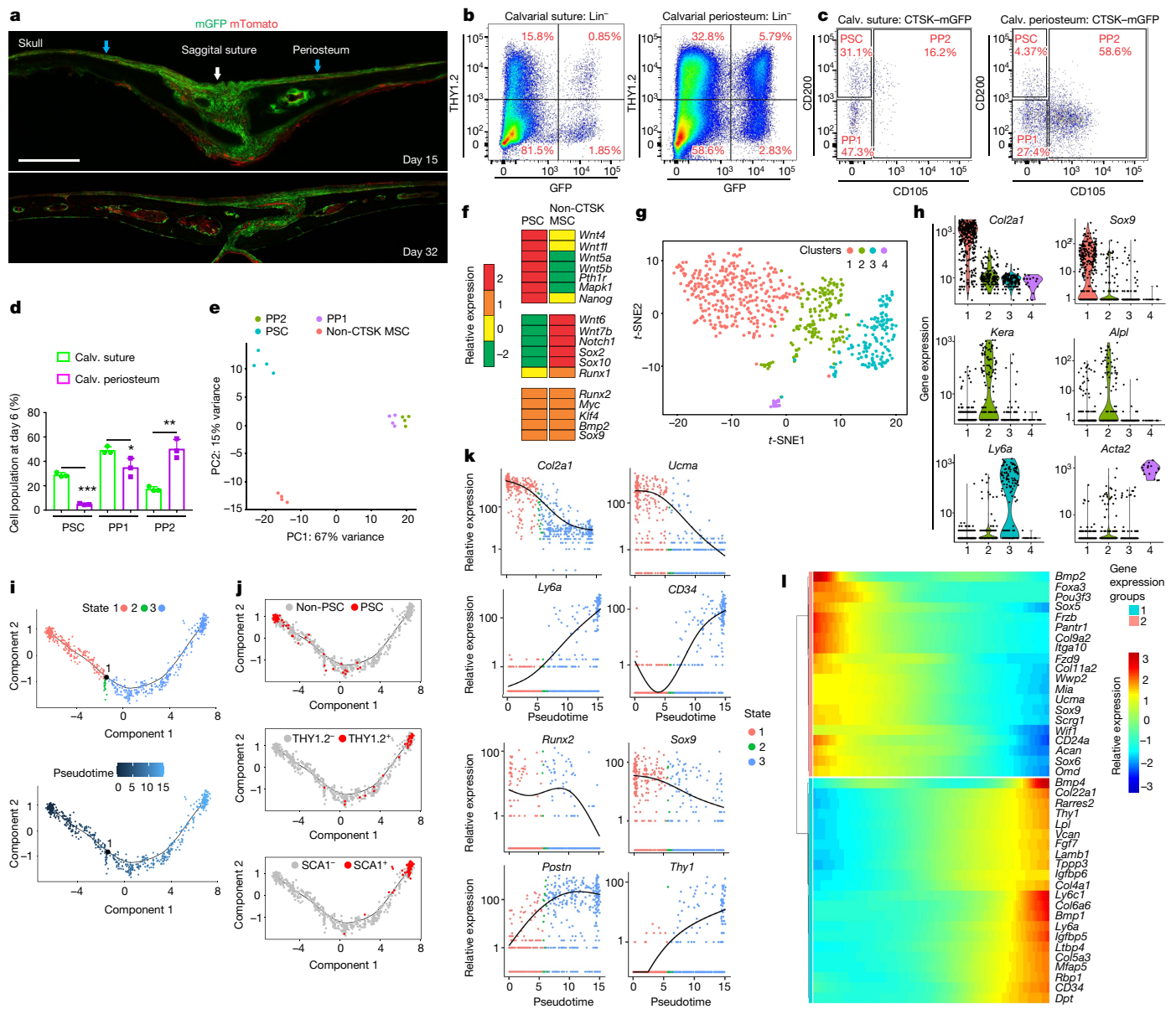


Fig. 3 | Periosteal stem cells in mouse calvarium and gene expression analysis of mouse femoral periosteal cells. **a**, CTSK-mGFP cells (green) in mouse calvarium at P15 (top) and P32 (bottom). White arrow, sagittal suture; blue arrow, calvarial periosteum; scale bar, 200 μ m. Images are representative of 3 independent experiments. **b**, THY1.2 expression in CTSK-mGFP cells from P6 suture (left) and calvarial periosteum (right). **c**, PSC, PP1 and PP2 in suture (left) and calvarial periosteum (right) at P6. Representative plots from 10 independent experiments. **d**, Relative proportion of PSC, PP1 and PP2 cell populations at day 6 in suture and calvarial periosteum. * $P=0.04$, ** $P=0.002$, *** $P=2.7 \times 10^{-5}$. Two-tailed Student's t -test. Data are mean \pm s.d.; $n=3$ independent experiments, 5 animals pooled per group. **e**, Principal component analysis of RNA-seq following FACS of PSC, PP1, PP2 and non-CTSK MSC populations from

P6 mouse femurs. $n=4$ mice per group. **f**, Heat map of gene expression in PSCs and non-CTSK MSCs. **g–k**, CTSK-mGFP⁺ mesenchymal cells ($n=658$) isolated by FACS from P6 mouse femur and analysed by CEL-Seq2. **g**, t -distributed stochastic neighbour embedding (t -SNE) of global gene expression. Cluster 1, 276 cells; cluster 2, 215 cells; cluster 3, 145 cells; cluster 4, 16 cells. **h**, Relative expression of *Col2a1*, *Sox9*, *Kera*, *Alpl*, *Ly6a* and *Acta2* among the four clusters in **g**. **i**, **j**, Monocle analysis of the CEL-Seq2 data. **i**, Single-cell trajectory obtained via unsupervised ordering of 658 CTSK-mGFP cells based on state (top) and pseudotime (bottom). **j**, Labelling on the differentiation trajectory of PSCs (top), THY1.2⁺ (middle) and SCA1⁺ (bottom) cells, determined by FACS analysis. **k**, Monocle analysis for relative gene expression versus pseudotime. **l**, Heat map of differentially expressed genes across pseudotime.

indicating that they are at the apex of the CTSK-mGFP differentiation hierarchy.

We further examined the specialization of PSCs for intramembranous bone formation at the calvarium, a well-known site of intramembranous bone formation. Consistent with observations that calvarial sutures contain progenitors that migrate to calvarial periosteum as they mature^{13,14}, cells with a PSC immunophenotype exist predominantly in the sutures (Fig. 3a, c, d, Extended Data Fig. 4d, e). By contrast, CTSK-mGFP⁺ PP1 and PP2 cells, as well as THY1.2⁺CD146⁺SCA1⁺ cells, were predominantly present in the

calvarial periosteum outside of the sutures (Fig. 3b–d, Extended Data Fig. 4f–j). Calvarial PSCs are functionally equivalent to long bone PSCs, being at the apex of their differentiation hierarchy during in vitro and in vivo assays, and possessing similar gene expression, the capacity to form tertiary mesospheres and bone organoids, and clonal multipotency (Extended Data Fig. 5). The few non-CTSK MSCs in the calvarium displayed low in vitro bone formation capacity (Extended Data Fig. 5d–f). Therefore, PSCs are present in the calvarial sutures, suggesting that PSCs orchestrate intramembranous ossification at multiple anatomic sites.

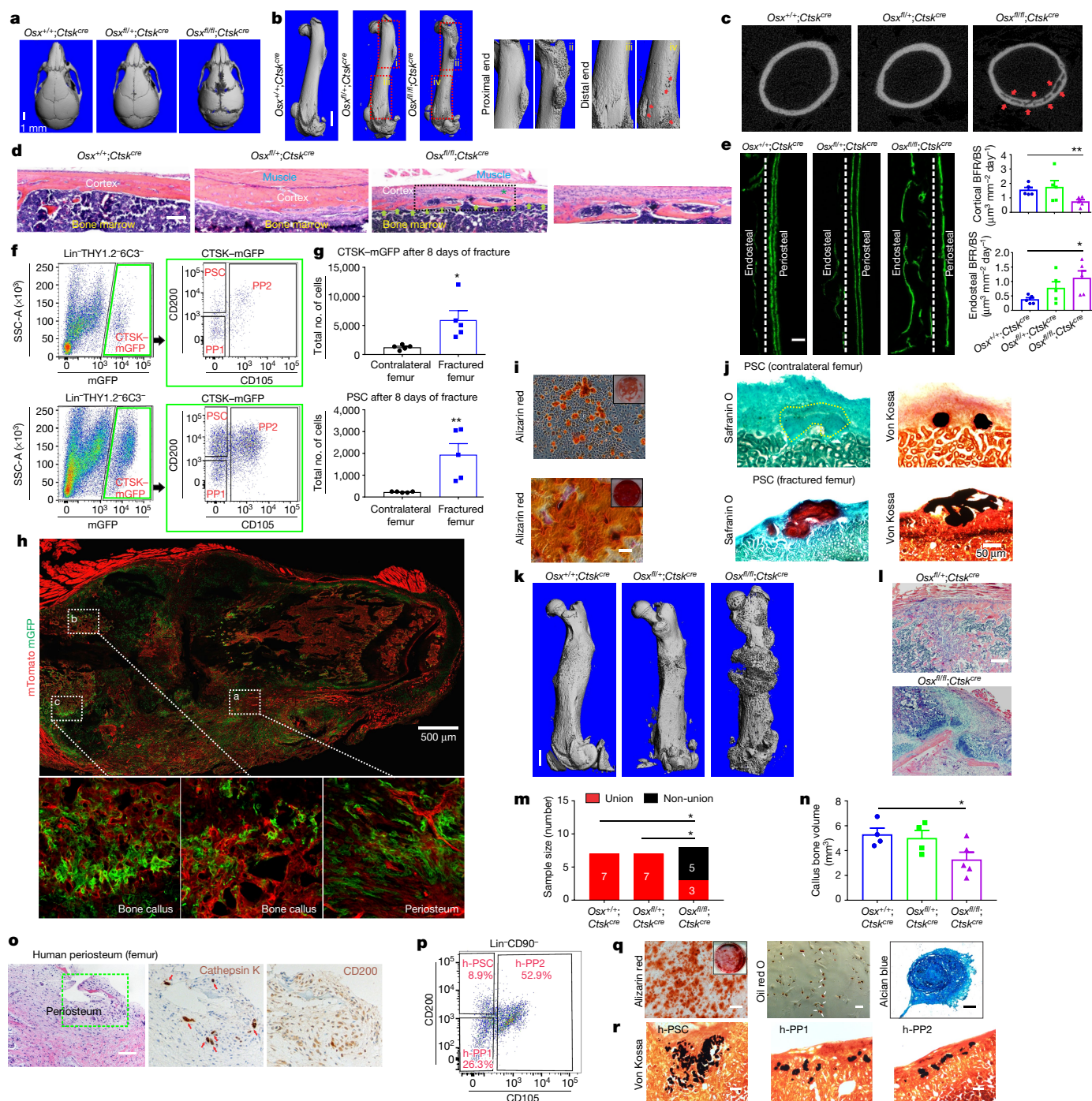


Fig. 4 | PSCs contribute to bone formation and fracture healing; human samples contain PSC-like cells. **a, b.** Micro-computed tomography (μ CT) of skull (**a**) and femur (**b**; smaller panels on right, enlarged view of outlined regions; arrows in iv indicate pitting that occurs on the surface of *Osx^{fl/yf};Ctsk^{cre}* mice) of 4-week-old mice. Scale bars, 1 mm. **c.** μ CT of femur cortex. Red arrows indicate cortical porosity. **d.** Left, haematoxylin and eosin staining of mouse periosteum; right, enlarged view of outlined region. Scale bar, 20 μ m. $n = 5$ animals per group. **e.** Calcein labelling of mouse femur (left) and bone formation rate/bone surface (BFR/BS; right) at 4 weeks old. Scale bar, 10 μ m. $*P = 0.021$, $**P = 0.007$; two-tailed Student's t -test. Data are mean \pm s.e.m., $n = 5$ animals per group, 5 independent experiments. **f, g.** FACS analysis of cells isolated from contralateral (top) and fractured femurs (bottom) (**f**, green outlines indicate parent/daughter gates) and cell counts (**g**). $*P = 0.018$; $**P = 0.009$ (callus tissue 8 days after fracture). Two-tailed Student's t -test. Data are mean \pm s.e.m., $n = 5$. **h.** *Ctsk^{cre};Rosa26^{mT/mG}* mouse femurs 9 days after fracture (top). Bottom panels show enlarged views of outlined regions. Scale bar, 500 μ m. **i.** Alizarin red staining of PSCs from fractured (bottom) or contralateral (top) femur. Scale bar, 200 μ m. **j.** Safranin O (left) and von Kossa (right)

staining 4 weeks after transplantation of PSCs from fractured (bottom) or contralateral femurs (top). **k–n**, Femur 3 weeks after fracture. **k**, μ CT. $n = 8$ mice per group. Scale bar, 1 mm. **l**, Haematoxylin and eosin staining of fracture callus. Scale, 200 μ m. **m**, Fracture non-union. $*P = 0.025$; Fisher's exact test. $n = 7$ ($Osx^{+/+}; Ctsk^{cre}$ and $Osx^{fl/+}; Ctsk^{cre}$), $n = 8$ ($Osx^{fl/fl}; Ctsk^{cre}$). **n**, Callus bone volume. $*P = 0.042$. Two-tailed Student's *t*-test. Data are mean \pm s.e.m. $n = 4$ ($Osx^{+/+}; Ctsk^{cre}$ and $Osx^{fl/+}; Ctsk^{cre}$), $n = 5$ ($Osx^{fl/fl}; Ctsk^{cre}$). **o–r**, Human femoral periosteum. **o**, Haematoxylin and eosin staining (left; scale bar, 200 μ m), and immunohistochemistry for cathepsin K (middle, red arrows) and CD200 (right). Middle and right panels show enlarged view of the region outlined in green in the left panel. **p**, FACS analysis. Representative plot from $n = 10$ experiments. **q**, Alizarin red (left, scale bar, 200 μ m; inset shows lower magnification), oil red O (white arrows; middle; scale bar, 50 μ m) and alcian blue (right; scale bar, 200 μ m) staining of cultured h-PSCs. **r**, Von Kossa staining (black) of bone organoids in xenografts of h-PSC (left), human PP1 (h-PP1) (middle) and human PP2 (h-PP2) (right). Scale bars, 20 μ m. Representative images from 3 (**h–j**, **l**, **q**) or 4 (**o**) independent samples.

Transcriptional analysis shows that PSCs from mouse femur are broadly different from both their PP1 and PP2 derivatives and other skeletal MSCs (Fig. 3e, Extended Data Fig. 6a, g). Both PSCs and non-CTSK-mGFP MSCs share expression of genes associated with mesenchymal stem or progenitor cells, including *Runx2* and *Sox9*, alongside stemness-associated genes such as *Myc*¹⁵ and *Klf4*¹⁶, whereas PSCs expressed higher levels of *Nanog* and *Wnt5a*¹⁷ (Fig. 3f). Notably, PSCs displayed increased per-cell bone formation capacity compared to PP1/PP2 cells after kidney capsule transplantation (Extended Data Fig. 6b–d). Calvarial PSCs also displayed similar patterns of gene expression as those observed in femoral PSCs (Extended Data Fig. 5g). In parallel, to empirically determine the populations present within the pool of femoral CTSK-mGFP⁺ Lin[−] cells, we performed single-cell RNA sequencing (RNA-seq) analysis using CEL-Seq2¹⁸. Mesenchymal CTSK-mGFP⁺ cells clustered into four groups: a group expressing progenitor/stem cell markers¹⁹ such as *Sox9* and *Col2a1*, a group expressing osteoblast markers such as *Bglap* and *Alpl*, a group expressing *Ly6a* (also known as *Sca1*) and a small group with high expression of *Acta2* (Fig. 3g, h, Extended Data Fig. 6e, f). Unsupervised construction of an inferred differentiation trajectory using Monocle²⁰ empirically identified a population corresponding to the PSCs obtained by FACS (Fig. 3i–j). Transcriptional markers of mesenchymal stem or progenitor cells were expressed in the early part of this differentiation trajectory containing PSCs (Fig. 3k, l). Therefore, combined index sorting and single-cell RNA-seq independently identify a discrete PSC population similar to that identified by FACS, and PSCs display transcriptional signatures of mesenchymal stem cells and are distinct from other CTSK-mGFP cell populations.

To evaluate the physiologic importance of PSC-derived osteoblasts to bone formation, we blocked the ability of PSCs to give rise to osteoblasts by conditionally deleting the osterix gene (*Osx*, also known as *Sp7*), which encodes a transcription factor essential for osteoblast differentiation^{21,22}, using *Ctsk*^{cre}. *Osx*^{fl/fl}; *Ctsk*^{cre} mice displayed hypomineralization of the calvarium, uneven periosteal surfaces, and extensive linear intra-cortical pores that gave a characteristic appearance of a double cortex (Fig. 4a–d, Extended Data Fig. 7a). Endosteal trabecular bone mass was not substantially altered (Extended Data Fig. 7d). Otherwise, bone morphology and growth plate architecture were intact in *Osx*^{fl/fl}; *Ctsk*^{cre} mice aside from a slight reduction in long bone length (Extended Data Fig. 7b, c). Bone formation rates were significantly reduced in the periosteum of *Osx*^{fl/fl}; *Ctsk*^{cre} mice, with a compensatory increase in endosteal bone formation (Fig. 4e, Extended Data Fig. 7e). TRAP staining showed no changes in osteoclast numbers²³ (Extended Data Fig. 7f, g). Consistent with histomorphometry results, the *in vivo* and *in vitro* osteogenic potential of both long bone and calvarial PSCs but not endosteal MSCs was reduced in *Osx*^{fl/fl}; *Ctsk*^{cre} mice (Extended Data Fig. 7h–j). Therefore, PSC-derived osteoblasts are necessary for periosteal bone formation and the establishment of normal cortical architecture.

Consistent with observations that the periosteum is necessary for fracture repair^{24,25}, PSC-derived osteoblasts expand in response to injury and are necessary for fracture healing. PSCs exhibited a greater expansion than non-CTSK MSCs in femurs 8 days after fracture (Fig. 4f, g, Extended Data Figs. 7k–p, 8, 9a–c). PSCs isolated from the fracture site displayed enhanced osteoblast differentiation capacity on a per-cell basis *in vitro* (Fig. 4i). Therefore, fracture markedly increases both the numbers and the osteoblast differentiation capacity of PSCs.

It is a longstanding apparent contradiction that periosteum undergoes intramembranous bone formation at baseline while also being necessary for endochondral fracture repair. We examined whether PSCs could explain this apparent contradiction by switching from a baseline intramembranous bone formation capacity to acquire endochondral bone formation capacity after fracture. No CTSK-mGFP labelling of chondrocytes is observed at baseline; however, CTSK-mGFP cells contributed approximately half of the chondrocytes present in the fracture callus (Extended Data Figs. 8a, 9d, e). PSCs isolated from the fracture callus mediated endochondral ossification after transplantation into the

kidney capsule (Fig. 4h, j, Extended Data Fig. 9f, g). Therefore, whereas there is a clear distinction between intramembranous-competent PSCs and endochondral-competent MSCs at baseline, injury can introduce plasticity or interconversion between these cell types. This offers an explanation for how the periosteum can contribute to the endochondral process of fracture repair despite being specialized for intramembranous formation at baseline.

Next, we investigated the functional contribution of PSC-derived osteoblasts to fracture healing in *Osx*^{fl/fl}; *Ctsk*^{cre} mice, which displayed markedly impaired fracture healing with increased rates of fracture non-union and a decrease in the total fracture callus bone volume (Fig. 4k–n). Histologic analysis of the fracture callus was consistent with defects in mineralization, showing decreased bone and increased cartilage in the callus (Extended Data Fig. 10a–g).

To establish whether a population analogous to PSCs exists in humans, human periosteal tissue from femur was sectioned and stained, showing that cells express cathepsin K and CD200 (Fig. 4o). Similarly, flow cytometry revealed the presence of a population (Lin[−]CD90⁺CD200⁺CD105[−]; hereafter h-PSCs) in human periosteum bearing an immunophenotype similar to that of mouse PSCs (Fig. 4p, Extended Data Fig. 10h–k). These human PSCs were multipotent as they underwent osteogenic, adipogenic and chondrogenic differentiation *in vitro* (Fig. 4q). When transplanted into the kidney capsule of immunocompromised mice, h-PSCs, h-PP1 and h-PP2 cells all mediated intramembranous bone formation (Fig. 4r, Extended Data Fig. 10l). Therefore, human periosteum contains a population analogous to PSCs.

In this study, we identified a stem cell that serves as a physiologic precursor of periosteal osteoblasts, and has a critical role in both fracture healing and modelling of the bone cortex. As the bone cortex is essential for the biomechanical resistance of bone to fracture^{26–29}, isolation of PSCs is likely to facilitate both the development of drugs that increase cortical bone thickness as well as cellular therapy for skeletal injury³⁰. Moreover, the existence of a discrete periosteal stem cell demonstrates that bone consists of multiple distinct pools of stem cell progenitors, which in turn enables the functional specialization of each of these stem cell types and their derivatives. In this regard, PSCs are specialized for intramembranous bone formation at baseline, providing a cellular basis for long-observed differences between intramembranous and endochondral bone formation. Lastly, the interconversion between PSCs and endochondral-competent populations suggests that such interconversions may occur more broadly and contribute to the cellular basis of skeletal pathology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0554-8>.

Received: 9 October 2017; Accepted: 13 August 2018;
Published online 24 September 2018.

- Mendez-Ferrer, S. et al. Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466**, 829–834 (2010).
- Zhou, B. O., Yue, R., Murphy, M. M., Peyer, J. G. & Morrison, S. J. Leptin-receptor-expressing mesenchymal stromal cells represent the main source of bone formed by adult bone marrow. *Cell Stem Cell* **15**, 154–168 (2014).
- Morikawa, S. et al. Prospective identification, isolation, and systemic transplantation of multipotent mesenchymal stem cells in murine bone marrow. *J. Exp. Med.* **206**, 2483–2496 (2009).
- Chan, C. K. et al. Identification and specification of the mouse skeletal stem cell. *Cell* **160**, 285–298 (2015).
- Worthley, D. L. et al. Gremlin 1 identifies a skeletal stem cell with bone, cartilage, and reticular stromal potential. *Cell* **160**, 269–284 (2015).
- Nakamura, T. et al. Estrogen prevents bone loss via estrogen receptor α and induction of Fas ligand in osteoclasts. *Cell* **130**, 811–823 (2007).
- Yang, W. et al. *Ptpn11* deletion in a novel progenitor causes metachondromatosis by inducing hedgehog signalling. *Nature* **499**, 491–495 (2013).
- Chan, C. K. et al. Clonal precursor of bone, cartilage, and hematopoietic niche stromal cells. *Proc. Natl Acad. Sci. USA* **110**, 12643–12648 (2013).

9. Sacchetti, B. et al. Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131**, 324–336 (2007).
10. Pinho, S. et al. PDGFR α and CD51 mark human Nestin⁺ sphere-forming mesenchymal stem cells capable of hematopoietic progenitor cell expansion. *J. Exp. Med.* **210**, 1351–1367 (2013).
11. Chan, C. K. et al. Endochondral ossification is required for haematopoietic stem-cell niche formation. *Nature* **457**, 490–494 (2009).
12. Bi, Y. et al. Identification of tendon stem/progenitor cells and the role of the extracellular matrix in their niche. *Nat. Med.* **13**, 1219–1227 (2007).
13. Maruyama, T., Jeong, J., Sheu, T. J. & Hsu, W. Stem cells of the suture mesenchyme in craniofacial bone development, repair and regeneration. *Nat. Commun.* **7**, 10526 (2016).
14. Zhao, H. et al. The suture provides a niche for mesenchymal stem cells of craniofacial bones. *Nat. Cell Biol.* **17**, 386–396 (2015).
15. Cartwright, P. et al. LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* **132**, 885–896 (2005).
16. Li, Y. et al. Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. *Blood* **105**, 635–637 (2005).
17. Yang, Y., Topol, L., Lee, H. & Wu, J. Wnt5a and Wnt5b exhibit distinct activities in coordinating chondrocyte proliferation and differentiation. *Development* **130**, 1003–1015 (2003).
18. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 77 (2016).
19. Ono, N., Ono, W., Nagasawa, T. & Kronenberg, H. M. A subset of chondrogenic cells provides early mesenchymal progenitors in growing bones. *Nat. Cell Biol.* **16**, 1157–1167 (2014).
20. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
21. Baek, W. Y., de Crombrughe, B. & Kim, J. E. Postnatally induced inactivation of osterix in osteoblasts results in the reduction of bone formation and maintenance. *Bone* **46**, 920–928 (2010).
22. Nakashima, K. et al. The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. *Cell* **108**, 17–29 (2002).
23. Aliprantis, A. O. et al. NFATc1 in mice represses osteoprotegerin during osteoclastogenesis and dissociates systemic osteopenia from inflammation in cherubism. *J. Clin. Invest.* **118**, 3775–3789 (2008).
24. Utvag, S. E., Grundnes, O. & Reikeraas, O. Effects of periosteal stripping on healing of segmental fractures in rats. *J. Orthop. Trauma* **10**, 279–284 (1996).
25. van Gestel, N. et al. Engineering vascularized bone: osteogenic and proangiogenic potential of murine periosteal cells. *Stem Cells* **30**, 2460–2471 (2012).
26. Allen, M. R., Hock, J. M. & Burr, D. B. Periosteum: biology, regulation, and response to osteoporosis therapies. *Bone* **35**, 1003–1012 (2004).
27. Ozaki, A., Tsunoda, M., Kinoshita, S. & Saura, R. Role of fracture hematoma and periosteum during fracture healing in rats: interaction of fracture hematoma and the periosteum in the initial step of the healing process. *J. Orthop. Sci.* **5**, 64–70 (2000).
28. Yukata, K. et al. Aging periosteal progenitor cells have reduced regenerative responsiveness to bone injury and to the anabolic actions of PTH 1–34 treatment. *Bone* **62**, 79–89 (2014).
29. Malizos, K. N. & Papatheodorou, L. K. The healing potential of the periosteum molecular aspects. *Injury* **36**, S13–S19 (2005).
30. Bianco, P. et al. The meaning, the sense and the significance: translating the science of mesenchymal stem cells into medicine. *Nat. Med.* **19**, 35–42 (2013).

Acknowledgements This project was funded by the Office of the Director of the NIH under award DP5OD021351 given to M.B.G. M.B.G. holds a Career Award for Medical Scientists from the Burroughs Wellcome Foundation and a Basil O'Connor Award from the March of Dimes. This content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health. We thank D. Ballon, B. He, S. Mukherjee and the Flow Cytometry Core, Genomics Resources Core, Optical Microscopy Core and the Citigroup Biomedical Imaging Core at Weill Cornell Medicine for their technical support, and B. Sleckman and T. Evans for insightful comments on the manuscript.

Reviewer information Nature thanks M. T. Longaker, M. Young and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.D. initiated the study and M.B.G. supervised the project. S.D. and M.B.G. conceived the project. S.D. designed, conducted experiments and analysed data. A.R.Y. performed all mouse surgeries. J.M. supervised flow cytometry. S.L., T.Z. and D.A.L. performed data analysis on bulk RNA-seq and single-cell RNA-seq. R.X., M.E. and J.-H.S. performed cell culture, RT-PCR, immunostaining and μ CT analysis. N.L., Y.L. and Y.S.Y. performed μ CT, histology and cryosectioning of samples. M.H., M.P.B. and J.H.H. provided access to human samples, helped with sample processing and supervised human studies. S.D. and M.B.G. prepared the manuscript. All authors read and approved the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0554-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0554-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. Cathepsin K-Cre (*Ctsk^{cre}*) mice³¹ were a gift from S. Kato (University of Tokyo). Floxed osterix (*Osx^{fl/fl}*) mice²² were a gift from B. de Crombrughe (University of Texas M. S. Anderson Cancer Center). *R26R^{confetti}* (Stock 017492), *Actb^{creER}* (Stock 004682), NOD/SCID/IL2R γ (NSG, Stock 005557) and *Rosa26^{mT/mG}* mice (Stock 007676)³² were purchased from Jackson Laboratories. All mice were maintained on a C57BL/6J background throughout the study. All animals were maintained in accordance with the NIH Guide for the Care and Use of Laboratory Animals and were handled according to protocols approved by the Weill Cornell Medical College subcommittee on animal care (IACUC). For all cellular transplantation experiments, the recipient animals for each specific cell population transferred were randomized after selecting a group of gender-appropriate mice. Generally, sample sizes were calculated on the assumption that a 30% difference in the parameters measured would be considered biologically significant with an estimate of sigma of 10–20% of the expected mean. Alpha and Beta were set to the standard values of 0.05 and 0.8, respectively.

Blinding. μ CT analysis, mesensphere calculation, histomorphometry, immunohistochemistry were performed by individuals that were blinded to the identity of the groups.

Cell line. Validation for single-cell sorting was initially done with RAW264.7 cells from ATCC (ATCC TIB-71), a commercially available mouse macrophage cell line. These cells were not independently authenticated since they were obtained directly from ATCC and were only used to validate single-cell sorting capabilities before experimentation. Cells were tested for mycoplasma using the Plasmotest Mycoplasma Detection Kit from InvivoGen (rep-pt2) and PCR analysis. All cells were confirmed to be free of mycoplasma.

Tamoxifen injection. *Actb^{creER}* activity was induced at postnatal day 3. Tamoxifen (100 mg/kg) (Sigma) was prepared in corn oil and injected intraperitoneally into nursing females for 3 consecutive days. Nursing females in control groups were simultaneously injected with corn oil only. The resulting litters were killed 14 days after induction and femurs were fixed for 4 h with 4% paraformaldehyde (PFA) and processed for imaging as described below.

Human samples. This project was approved by the IRB committee of Memorial Sloan Kettering Cancer Center (MSKCC) (IRB No. 97-094). HIPAA authorization was included in the IRB protocol and all ethical guidelines conform to the 2008 Helsinki declaration. Patient consent was obtained and all identifying patient information was scrubbed at the time of sample procurement.

Isolation of mesenchymal cells. Skeletal tissue from postnatal day 7, 15 or 32 mice was subjected to both mechanical and enzymatic digestion. The harvested tissue was minced using razor blades and digested for up to 1 h with Collagenase P (1 mg/ml; Roche, cat. 11213857001) and Dispase II (2 mg/ml; Roche, cat. 04942078001) digestion buffer at 37°C with agitation. Medium containing 2% serum was added to the digest and the tubes were centrifuged to pellet cells. The supernatant was removed, and the pellet was resuspended in DNase I (2 units/ml) solution and briefly incubated for 5 min at 37°C. Medium was added to the tube and the digested tissue was resuspended thoroughly by pipetting and then filtered through 70- μ m nylon mesh. Tubes were centrifuged and the resulting cell pellet was subjected to FACS. Microdissection of mouse periosteum, calvarial suture and calvarial periosteum was performed under a dissecting microscope and the resulting tissue samples were subjected to the same digestion protocol. For human samples, tissue was treated with Pronase (1 mg/ml; Roche, cat. 10165921001) for 20 min at 37°C under agitation and spun down at 1500 r.p.m. for 10 min before collagenase treatment as above.

Fluorescence assisted cell sorting (FACS). Cells were washed twice with ice-cold FACS buffer (2% FBS + 1 mM EDTA in PBS), incubated with blocking buffer (1:100 dilution; BD bioscience 553142 for mouse and 564765 for human) for 15 min at 4°C. Primary antibody dilutions were prepared in brilliant stain buffer (BD bioscience 563794). Cells were incubated in the dark for 1 h on ice with primary antibody solution, washed 2–3 times with FACS buffer, and incubated with secondary antibody solution for 20 min. Cells were then washed several times and resuspended in FACS buffer with DAPI (4',6-diamidino-2-phenylindole; 1 μ g/ml; Invitrogen D1306). FACS was performed using a Becton Dickinson Aria II equipped with 5 lasers (BD bioscience). Beads (eBioscience 01-1111) were used to set initial compensation. Fluorescence minus one (FMO) controls were used for additional compensation and to assess background levels for each stain. Gates were drawn as determined by internal FMO controls to separate positive and negative populations for each cell surface marker. Typically, 1–2 million events were recorded for each FACS analysis, and the data was analysed using FlowJo (v.10.1). Calculations were made in Excel and bar graphs were generated in GraphPad Prism. To better depict the FACS gating strategy, we have made changes throughout our figures, using colour-coded boxes to illustrate parent/daughter gates. Additionally Extended Data Fig. 2d has been provided to show the full details of the gating strategies used.

FACS antibodies. Antibodies for FACS of human samples included CD235a (clone HIR2, eBioscience), CD31 (clone WM-59, eBioscience), CD45 (BD Bioscience),

CD90 (clone 5E10, BD Bioscience), CD140 α (clone aR1, BD Bioscience), CD105 (clone 266, BD Bioscience), CD200 (clone MRC OX-104, BD Bioscience), CD146 (clone P1H12, BD Bioscience), CD140 α (clone α R1, BD Bioscience), CD49f (clone GoH3, BD Bioscience), CD51 (clone NKI-M9, Biolegend) and CD295 (clone 52263, BD Bioscience). Antibodies for FACS of mouse samples included CD45 (clone 30-F11, BD Bioscience), CD31 (MEC13.3, BD Bioscience), Ter119 (clone Ter119, BD Bioscience), CD90.2 (clone 53-2.1, BD Bioscience), BP-1 (clone 6C3, eBioscience), CD200 (clone OX-90, BD Bioscience), CD105 (clone MJ7/18, Biolegend), CD146 (clone ME-9F1, BD Bioscience), CD140 α (clone APA5, Biolegend), CD49f (clone GoH3, BD Bioscience), CD51 (clone RMV-7, BD Bioscience), Ly6A/E (clone D7, BD Bioscience), leptinR (R&D systems), CD61 (clone 2C9.G2, BD Bioscience), streptavidin eFluor 710 (eBioscience) and streptavidin BUV737 (BD Bioscience).

Cell culture. Sorted cells were cultured both under hypoxic (2% O₂, 10% CO₂, 88% N₂) and non-hypoxic conditions (20% O₂). In vitro differentiation of FACS-isolated populations was conducted under hypoxic conditions (Fig. 2d, Extended Data Figs. 2i, 3h, 5k–m). Clonal multipotency of sorted cell populations was evaluated under non-hypoxic conditions (Figs. 2e, 4q, Extended Data Figs. 3d, 5c). Mesensphere formation capacity was evaluated under non-hypoxic conditions (Fig. 2a, Extended Data Figs. 3f, 5a).

Mouse primary cells were grown in complete mesencult medium (basal medium, 05501, Stem Cell Technologies) with stimulatory supplements (05502, 05500, Stem Cell Technologies). After initial cell plating, cells were left undisturbed and allowed to grow at 37°C under humidified conditions for a week. Half of the medium was replaced every 7 days. Cells were passaged once they were 60–70% confluent with Stem Pro Accutase solution (Gibco, A11105-01). Sorted human cells were cultured under conditions similar to those described above using a commercial medium preparation (basal medium, 05401, Stem Cell Technologies) with stimulatory supplements (05402, Stem Cell Technologies).

Mesensphere assays. For mesensphere assays, FACS-isolated single cells were plated at a density of 100 cells per cm² and allowed to grow in ultra-low adherence culture dishes (Stem Cell Technologies, 27145). Plates were incubated at 37°C with 5% CO₂ and left undisturbed for a week. Half of the medium was replaced every 7 days. Mesenspheres were dissociated into single cells using Accutase solution (Gibco, A11105-01) and were subsequently re-plated to generate secondary and tertiary mesenspheres.

Clonal differentiation assay. The differentiation potential of 10 colonies derived from single FACS-isolated PSCs was examined (Fig. 2e). In brief, single FACS-sorted cells were plated at a density of 100 cells per 10-cm dish and allowed to form individual colonies. Initial dispersion of the plated cells as single cells was confirmed by light microscopy. Each of the selected colonies was extracted using a cloning cylinder. The extracted cells were regrown for 3–4 days in 12-well plates and then allowed to differentiate under both osteogenic and adipogenic conditions as described below. A similar approach was used to analyse clonal differentiation of FACS-sorted calvarial PSCs (Extended Data Fig. 5c).

Osteogenic differentiation and alizarin red staining. Sorted cells were expanded and then allowed to differentiate using osteogenic differentiation medium (Stem Pro Osteogenesis Differentiation kit, A10072-01) for 19–28 days. Medium was changed every 2 days. At the end of this period, cells were washed with cold PBS and fixed with 70% ethanol for 15 min on ice. Cells were washed with distilled water and stained with alizarin red solution for 2–3 min. Cells were then washed thoroughly with water and air dried before microscopic visualization.

Adipogenic differentiation and oil red O staining. Adipogenic differentiation was similarly conducted with sorted cells as described above. In brief, cells were allowed to differentiate in adipogenic differentiation medium (Stem Pro Adipogenesis Differentiation kit, A10070-01). Fresh medium was added every 2–3 days for a total of 14–20 days. Medium was removed, and cells were washed with PBS and fixed with 4% PFA for 30 min at room temperature. Cells were rinsed again with PBS and stained for 30 min with oil red O working solution (3:2 dilution with water). Cells were then observed under a light microscope after 4–5 washes with PBS.

Chondrogenic differentiation and alcian blue staining. Micromass cultures were generated by seeding 1 \times 10⁵ cells in 5–10 μ l of basal medium. Cells were allowed to adhere to a glass slide for 2 h under highly humidified conditions at 37°C. Chondrogenic differentiation medium (Stem Pro Chondrogenesis Differentiation kit, A10071-01) was slowly added and cells were incubated at 37°C with 5% CO₂. Cultures were re-fed every 2–3 days and allowed to differentiate for a minimum of 14 days. At the end point of this study, medium was removed, and cultures were washed with PBS and fixed with 4% PFA for 30 min. Cultures were stained for 30 min with 1% alcian blue solution, washed three times with 0.1 N HCl and then with PBS.

Separately, pellet culture was conducted with 250 \times 10³ cells seeded in 15-ml polypropylene tubes. The pellet was cultured in chondrogenic differentiation medium, which was changed every 2–3 days for a period of 2 weeks. The pellet was harvested, fixed with 4% PFA and embedded in OCT with 15% sucrose

solution. Sections, 10 µm in thickness, were cut and alcian blue staining was performed as described.

Surgical procedures. All surgical procedures were performed under isoflurane (1–4%) anaesthesia. Surgical sites were sterilized using a betadine/iodide/isopropanol prep after hair removal using a clipper with a no. 40 blade and a depilatory cream (Nair). After surgery, the visceral lining or muscle was sutured with absorbable Ethicon vicryl sutures (VWR, 95057-014) before closing the skin with wound clips that were removed 2 weeks post-operatively. Animals received intraperitoneal buprenex (0.5 mg/kg) and oral meloxicam (2.0 mg/kg) as analgesia before surgery and for every 24 h post-surgery for 3 days. All surgical procedures are approved by the institutional animal care and use committee at Weill Cornell Medical College.

Kidney capsule transplantation model. In brief, 8–10-week-old male mice were anaesthetized and shaved on the left flank and abdomen before sterilization of the surgical site. The kidney was externalized through a 1-cm incision and a 2-mm pocket was made in the renal capsule. A 5-µl Matrigel plug (Corning, 356231) containing 8,000–10,000 cells was implanted underneath the capsule and the hole was sealed using a cauterizer before replacing the kidney back into the body cavity. Recipients for these experiments were syngenic with donors. Moreover, to avoid potential immunogenicity of GFP itself, all transplantation studies of CTSK-mGFP⁺ cells were conducted in *mTmG* hosts that have baseline immunologic tolerance to GFP variants. Animals were euthanized by CO₂ after 6 weeks. After death, kidneys were fixed with 4% PFA for 5 h and bone formation was detected by µCT. Samples were subjected to infiltration, embedding and sectioning as described below. Haematoxylin and eosin and Von Kossa staining were performed following previously described protocols³³. Standard safranin O stain and alizarin red staining was performed to detect cartilage and bone components.

Altogether, over 30 individual mouse transplantation procedures were conducted across 5 separate days of experimentation to study the intramembranous versus endochondral differentiation preferences of PSCs versus non-CTSK MSCs. **Mammary fat pad transplantation model.** To facilitate distinguishing host cells from CTSK-mGFP⁺ transplanted cells, 3–4-week-old female *mTmG* mice were used, where all host cells were mTomato⁺. These mice were anaesthetized and their lower abdomen was shaved before sterilization of the surgical site. A 1–1.5-cm longitudinal incision was made along the midline of the body wall below the sternum and just above the pelvic region. A second 0.5-cm cut was made from the bottom of the midline incision towards the hip of the right hind limb. The skin was carefully pulled away from the peritoneum to expose the inguinal mammary gland. A 2-mm pocket was created in the mammary fat pad just below the lymph node and a 5-µl Matrigel plug with the desired number of cells was implanted into the pocket. The pocket was sealed by clamping the edges of the cut fat pad together with a toothed forcep. Wound clips were used to close the skin incisions. Animals were euthanized by CO₂ at 2.5 weeks post-surgery, and the mammary fat pad was extracted, minced with razor blades and digested at 37 °C in digestion buffer (Collagenase P (1 mg/ml) and Dispase II (2 mg/ml) in 2% serum medium) for 30–45 min. Medium was added and the tubes were spun for 10 min at 1,500 r.p.m. The supernatant was carefully removed and DNase I solution was added for 5 min without disturbing the pellet. The reaction was terminated by diluting the digestion solution with medium. Tubes were again spun for 10 min and the resulting pellet was prepared for FACS analysis.

Femur fracture model. In brief, after anaesthesia, an incision was made above the right anterolateral femur after sterilization of the surgical site. The femur and patella were then exposed and a 27-gauge syringe needle was inserted parallel with the long axis of the femur through the patellar groove into the marrow cavity. The needle was then removed and a single cut in the mid-diaphysis of the femur was made using a Dremel saw with a diamond thin cutting wheel (VWR, 100230-724). A blunt 25-gauge needle was then reinserted into the marrow space through the hole made in the femur to stabilize the fracture. This needle was then trimmed so it would not project into stifle joint space. Muscle was then placed over the injury site and stitched with absorbable sutures before closing the skin with wound clips. We performed femur fracture in 6-week-old *Ctsk^{cre};mTmG* mice (Fig. 4f–h) and 4-week-old *Osx^{fl/fl};Ctsk^{cre}* mice (Fig. 4k). Mice were killed by CO₂ asphyxiation after 3 weeks.

Sample preparation for cryo-sectioning. Freshly extracted mouse samples were fixed with 4% PFA for 4 h at 4 °C. Samples were washed with PBS and decalcified with 0.5 M EDTA for 1–5 days depending on the age of the sample. Samples were incubated with infiltration solution (20% sucrose + 2% polyvinylpyrrolidone in PBS) with rocking until they sank to the bottom of the tube. Embedding was performed with OCT + 15% sucrose and samples were preserved at –80 °C. Sections, 10–20 µm in thickness, were cut using a Leica cryostat.

Immunohistochemistry. Frozen samples were thawed at room temperature and rehydrated with PBS, permeabilized with PBS + 0.3% Triton X-100 for 15 min, and blocked for 1 h with 5% donkey serum in PBS (blocking buffer). Dilutions of primary antibodies were freshly prepared in blocking buffer. Samples were incubated overnight with primary antibodies at 4 °C, then washed three times with

PBS. Secondary antibodies (1:2,000 dilution) were added to the sample for 1–2 h, followed by washing three times with PBS. DAPI (300 nM) was added for 5–10 min and the samples were mounted with antifade mounting solution (Life technologies, P36970). Imaging was performed with a Zeiss LSM 880 with Airyscan high-resolution-detector confocal microscope. For immunohistochemistry of mesenspheres, mesenspheres were extracted and allowed to adhere on glass chamber slides for 24 h. Spheres were fixed with 4% PFA for 15 min at room temperature, washed twice with PBS, permeabilized with PBS + 0.5% Triton X-100 for 10 min and blocked for 1–2 h with blocking buffer (PBS + 2% BSA + 10% horse serum + 0.5% Triton X-100). Primary antibody dilutions were prepared in blocking buffer, and added to the spheres followed by overnight incubation at 4 °C. Samples were washed three times with PBS and incubated with secondary antibody for 1 h and DAPI solution for 10 min. Samples were subsequently processed for imaging as described above.

Primary antibodies for immunohistochemistry. Primary antibodies used were specific for collagen type I (Abcam, ab34710, 1:100 dilution), collagen type II (Millipore, MAB8887, 1:100 dilution), murine CD200 (Abcam, ab33734, 1:100 dilution), nestin (Abcam, ab11306, 1:200 dilution), gremlin 1 (Abcam, ab189267, 1:50 dilution), COMP (Abcam, ab74524, 1:50 dilution), Aggrecan (Abcam, ab3778, 1:100 dilution), THY1.2 (Invitrogen, 14-0902-82, 1:50 dilution), 6C3 (Invitrogen, 14-5891-82 1:50 dilution), CD105 (Abcam, ab107595, 1:100 dilution), Runx2 (Abcam, ab76956, 1:200 dilution), Alpl (Abcam, ab108337, 1:100 dilution), osteocalcin (Abcam, ab93876, 1:100 dilution), CD146 (Abcam, ab75769, 1:100 dilution), CD140α (Abcam, ab96569, 1:100 dilution), CD200 (Abcam, ab203887, 1:200 dilution), tartrate resistant acid phosphatase (TRAP) (Abcam, ab185716, 1:50 dilution), and cathepsin K (Abcam, ab19027, 1:200 dilution).

Calcein labelling. Calcein stock solution (2.5 mg/ml) was freshly prepared in calcein buffer (ddH₂O + 2% NaHCO₃ + 150 mM NaCl). Four-week-old *Osx^{fl/fl};Ctsk^{cre}* mice from the same litter were given subcutaneous calcein injection (10 mg/kg) followed by a second injection after 2 days. Mice were euthanized after 24 h, fixed overnight in 10% formalin solution and then stored in 70% ethanol. Non-decalcified femur sections were prepared for plastic embedding according to a previously published protocol³³.

Histomorphometry. Mice underwent dual calcein labelling, and undecalcified femur sections were prepared for analysis and stained for TRAP as previously described³³. Dynamic histomorphometry parameters were measured using the Osteomeasure Analysis system (Osteometrics) and reported following standard nomenclature³⁴. Mineral apposition rate (MAR; µm day^{−1}), bone formation rate/bone surface (BFR/BS; µm³ mm^{−2} day^{−1}), bone volume (BV; mm³), bone volume/total volume (BV/TV), osteoclast number/bone perimeter (No. Oc/B. Pm) were determined.

µCT and paraffin embedding. µCT was conducted following the previously described parameters³⁵ on a Scanco Medical Micro-CT35 system at the Citigroup Biomedical Imaging Core. µCT operators and evaluators were blinded to the experimental groups under analysis. Human periosteal samples were fixed overnight with 4% PFA at 4 °C. Samples were decalcified with 0.5 M EDTA for 5–6 days and then subjected to paraffin embedding. The Translational Research Pathology Core at Weill Cornell performed paraffin embedding, sectioning and staining (haematoxylin and eosin, CD200 and cathepsin K stain for human samples) following a previously published protocol³⁵.

Bulk RNA-seq. Bulk RNA-seq was performed on sorted PSC, PP1, PP2, and non-CTSK MSC populations isolated from 6-day-old mouse femurs. cDNA libraries were generated using the Illumina TruSeq RNA Sample Preparation kit and sequenced on HiSeq4000 sequencer. Tophat2 was used to align raw sequencing reads to the mm10 mouse reference genome. Differential expression analysis was performed using DESeq2 package. The heat map was generated using heatmap.2 in the R gplots package, where the expression values were normalized per gene over all samples³⁶. For each gene, the mean and standard deviation (s.d.) of expression over all samples was calculated, and the expression value underwent linear transformation using the formula (RPKM – mean)/s.d. For 16 bulk RNA samples, the average raw reads per sample was 46.2 million.

Single-cell RNA-seq using CEL-Seq2. Validation for single-cell sorting was initially done with RAW264.7 cells using 384-well plates. Single-cell sorting was confirmed through observation of growth of a single colony from the sorted cells (Extended Data Fig. 6h). We found that cells were sorted into wells at 93% efficiency and the doublets were detected in less than 2% of cases (Extended Data Fig. 6i).

For the experiment, CTSK-mGFP⁺ cells isolated from a 6-day-old mouse femur were sorted by FACS (Becton-Dickinson Influx) and plated into individual wells of a 384-well plate pre-loaded with unique barcoded reverse-transcription primers. After sorting, cells were snap-frozen on dry ice before being submitted to the New York Genome Center (NYGC) for cDNA synthesis and library preparation. cDNA synthesis was performed in individual wells using a template-switching mechanism followed by PCR amplification. All barcoded amplified cDNA was then pooled and run on a pico-green assay and fragment analyzer to evaluate sample concentration

and quality, respectively. Libraries were generated from 300 pg of pooled cDNA using the Nextera XT library preparation kit (Illumina). The library was then sequenced with custom sequencing primers across two lanes of a 25 bp × 75 bp rapid run (HiSeq 2500 instrument). Read 1 was composed of the cell barcode and unique molecular identifier (UMI), whereas read 2 was used to map the transcript to the mouse genome. After sequencing, the alignment and analysis was done using the NYGC re-tooled pipeline based on Drop-seq Core tools³⁷ (<http://mccarrolllab.org/dropseq/> which provided basic quality control and expression analysis. The delivered data contained FASTQ, BAM files and an expression matrix.

Seurat analysis. Seurat v.2.0.1 and R v.3.4.1 were used for analysis of single-cell RNA-seq data³⁸. Cells expressing at least 200 but not more than 5,500 genes were retained for analysis. Genes expressed in less than 3 cells were not taken into consideration. A total of 658 single cells were analysed for the total number of UMIs per cell, percentage of mitochondrial gene expression, percentage of ribosomal gene expression, and cell cycle. Initial clustering analysis showed an even distribution of mitochondrial gene content across clusters, so this was not considered further (data not shown). Cell cycle analysis was performed using a previously published approach³⁹. Regression was performed to remove effects associated with cell cycle state, the number of UMIs and ribosomal gene content. Clustering was performed using the first 12 principal components and clusters were visualized using *t*-SNE projection.

Monocle analysis. Cell trajectory analysis was performed by using Monocle on CTSK-mGFP cells. Monocle v.2.4.0 and R v.3.4.0 were used to conduct analysis on the single-cell sequencing data. Monocle analysis^{40–42} was performed using negbinomial.size function and the data was loaded using sparseMatrix. Quality control was performed by filtering low-quality cells including dead cells, empty wells in the plate and doublets. The mRNA threshold was defined and cells with mRNA content between 35,000 through 295,000 were analysed (Extended Data Fig. 6j). The same CEL-Seq2 dataset as used above was loaded for analysis (Extended Data Fig. 6k). Cell trajectory was obtained using genes selected by a principal-component-analysis-based method for cell ordering according to previously described methods^{40–42}. Regression was performed to remove effects associated with cell cycle state and ribosomal gene content. A list of differentially expressed genes that change as a function of pseudotime was obtained using fullModelFormulaStr. Genes that were significant at 10% false discovery rate were selected. Heat maps to visualize pseudotime-dependent genes were generated by using a subset of the significant genes.

Statistical analyses. All data are shown as the mean ± s.d. or mean ± s.e.m. as indicated. Where applicable, we first performed the Shapiro–Wilk normality test to check normality. For comparisons between two groups, a two-tailed, unpaired

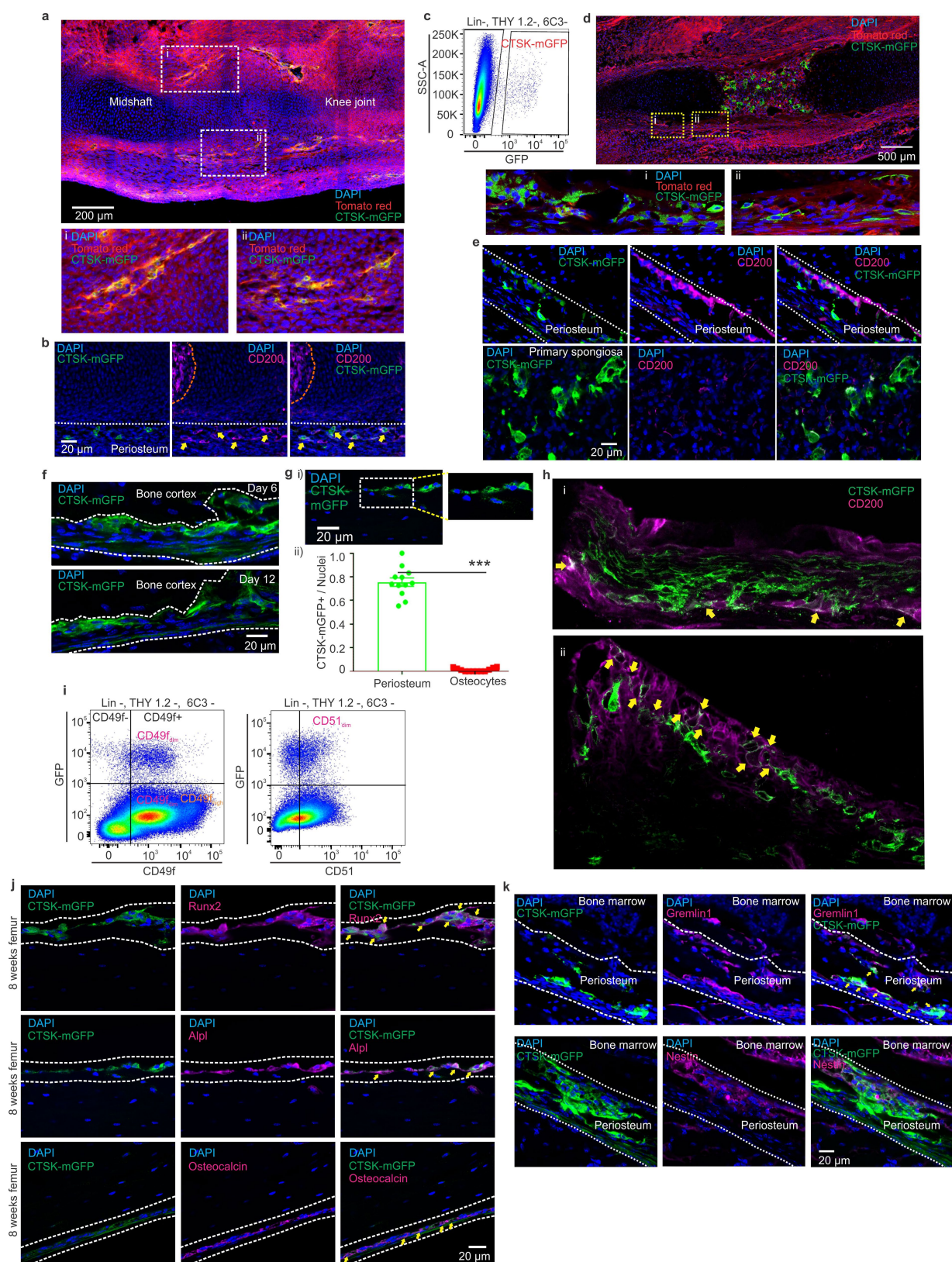
Student's *t*-test was used. If normality tests failed, Mann–Whitney *U*-tests were used. For comparisons of three or more groups, one-way ANOVA was used if normality tests passed, followed by Tukey's multiple comparison test for all pairs of groups. If normality tests failed, Kruskal–Wallis test was performed, followed by Dunn's multiple comparison test. GraphPad PRISM v.6.0a was used for statistical analysis. *P* < 0.05 was considered statistically significant.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-seq and single-cell RNA-seq data have been deposited at the Gene Expression Omnibus under accession numbers GSE106237, linked to the subseries GSE106235 and GSE106236. Microscopy images are available from the corresponding author upon reasonable request.

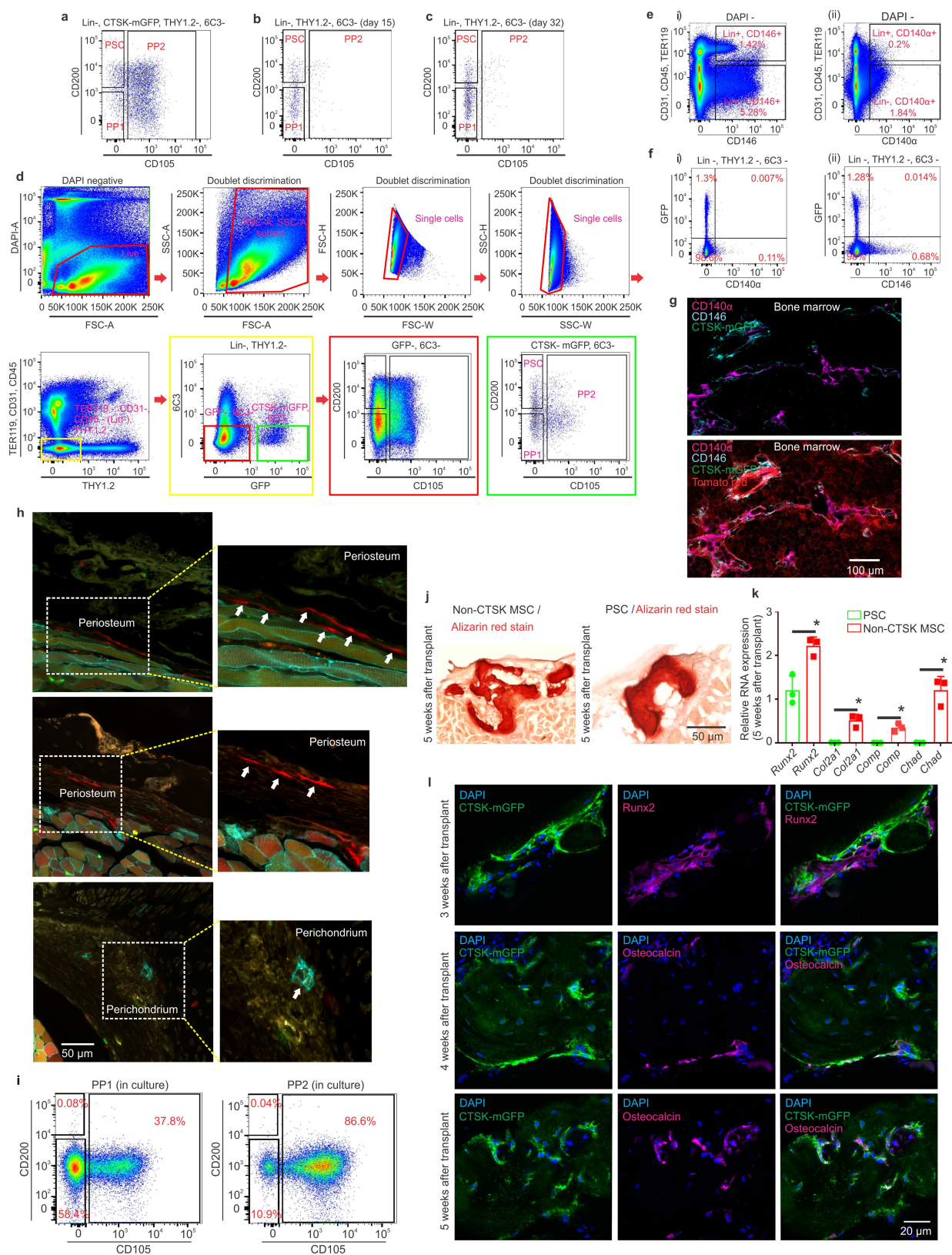
31. Nakamura, T. et al. Estrogen prevents bone loss via estrogen receptor α and induction of Fas ligand in osteoclasts. *Cell* **130**, 811–823 (2007).
32. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
33. Fukuda, T. et al. Sema3A regulates bone-mass accrual through sensory innervations. *Nature* **497**, 490–493 (2013).
34. Dempster, D. W. et al. Standardized nomenclature, symbols, and units for bone histomorphometry: a 2012 update of the report of the ASBMR Histomorphometry Nomenclature Committee. *J. Bone Miner. Res.* **28**, 2–17 (2013).
35. Greenblatt, M. B. et al. CHMP5 controls bone turnover rates by dampening NF- κ B activity in osteoclasts. *J. Exp. Med.* **212**, 1283–1301 (2015).
36. Crespo, M. et al. Colonic organoids derived from human induced pluripotent stem cells for modeling colorectal cancer and drug testing. *Nat. Med.* **23**, 878–884 (2017).
37. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
38. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
39. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
40. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
41. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
42. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Analysis of CTSK-mGFP cells in mouse femur. **a**, CTSK-mGFP mesenchymal cells (green) were visualized in the mouse long bones at E14.5. Scale bar, 200 μm . Enlarged images of areas marked by the dotted white boxes are provided in **i** and **ii**. **b**, Immunostaining for CD200 (magenta) confirmed co-localization (shown by yellow arrows) with CTSK-mGFP cells (green) in the periosteum. A separate pool of CD200⁺ cells are detected at the future primary ossification site (marked by dotted orange line). Scale bar, 20 μm . Images in **a** and **b** are representative of 3 independent experiments. **c**, CTSK-mGFP mesenchymal cells in the long bones of mice were detected by FACS at E16.5. **d**, Visualization of CTSK-mGFP cells (green) in mouse long bones at E16.5. Scale bar, 500 μm . An enlarged view of the areas marked by dotted yellow boxes are shown in **i** and **ii**. CTSK-mGFP cells (green) were detected in the mouse periosteum (**i** and **ii**). **e**, CD200 (magenta) immunostaining confirmed co-localization with CTSK-mGFP cells (green) in the periosteum (top panels). CTSK-mGFP cells in the primary spongiosa morphologically consistent with osteoclasts stained negative for CD200 (bottom panels). Scale bar, 20 μm . Images in **c–e** are representative of 3 independent experiments. **f**, Visualization of CTSK-mGFP cells (green) in the periosteum (dotted white line) of mouse femur

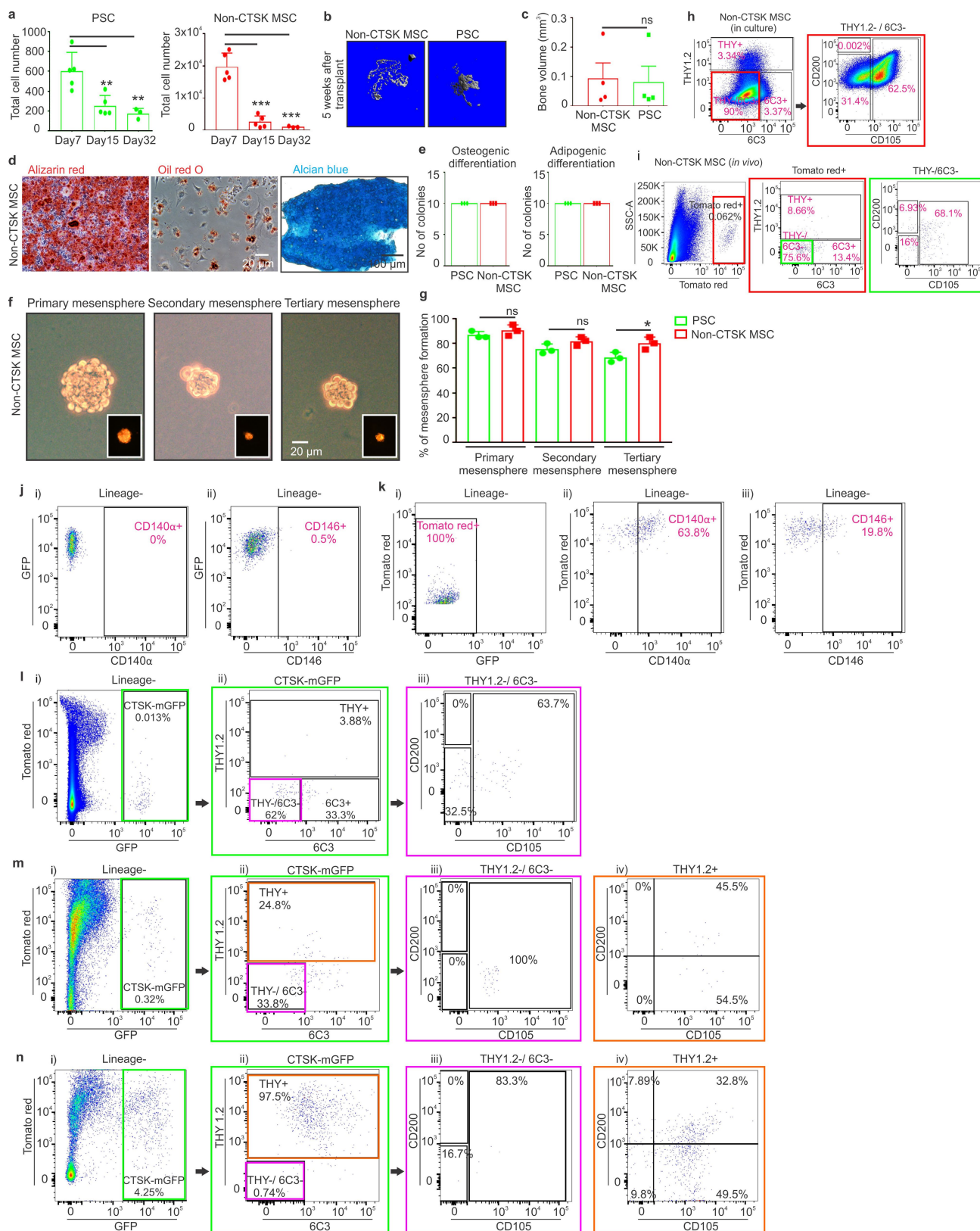
at postnatal day 6 (top) and 12 (bottom). Scale bar, 20 μm . **g**, CTSK-mGFP visualization shows rare mGFP⁺ osteocytes, an enlarged view of the dotted white box is provided (**i**). Scale bar, 20 μm . Image representative of 3 independent experiments. Quantification of total CTSK-mGFP-labelled periosteal cells and mGFP-labelled osteocytes in the mouse femur (**ii**). *** $P = 6.95 \times 10^{-16}$; two-tailed Student's *t*-test. Data are mean \pm s.e.m., $n = 12$ distinct areas of periosteum from 3 independent experiments. **h**, An enlarged view from Fig. 1e. Representative images from 3 independent experiments. **i**, FACS plots showing expression of CD49f (left) and CD51 (right) in CTSK-mGFP cells isolated from long bones of 7-day-old mice. Representative plot from 5 independent experiments. **j**, Femurs from 8-week-old *Ctsk^{cre}* mice were immunostained for Runx2 (magenta, top), alkaline phosphatase (ALPL) (magenta, middle) and osteocalcin (magenta, bottom). Co-localization is shown by yellow arrows. Scale bar, 20 μm . Representative images from 3 independent experiments. **k**, Femurs of 12-day-old *Ctsk^{cre}* mice were immunostained for gremlin 1 (magenta, top) and nestin (magenta, bottom). Dotted white line indicates periosteum. Scale bar, 20 μm . Representative images from 3 independent experiments.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | FACS analysis of microdissected periosteal tissue and characterization of PSCs. **a**, Flow cytometry of CTSK-mGFP cells microdissected from the periosteum of P7 mouse long bones, showing the distribution of PSC, PP1 and PP2 cells. **b**, **c**, Flow cytometry showing the distribution of PSCs, PP1 and PP2 cells in mouse long bones at day 15 (**b**) and day 32 (**c**). Plots in **a–c** are representative of results from 10 independent experiments. **d**, Schematic representation of the strategy used for FACS analysis of periosteal PSC, PP1 and PP2 cell populations. **e**, FACS plot showing the distribution of CD146 (i) and CD140 α (ii) expression in bone marrow stromal cells. **f**, FACS plots displaying the distribution of CD140 α (i) and CD146 (ii) in mouse periosteum obtained through periosteal microdissection. Plots in **e** and **f** are representative of results from 5 independent experiments. **g**, Mouse bone marrow immunostained for CD146 (cyan) and CD140 α (magenta). Scale bar, 100 μ m. Representative images from 3 independent experiments. **h**, Clonogenic cells detected in the periosteum (top and middle, white arrows) and perichondrium region (bottom, white arrows) of mouse femur

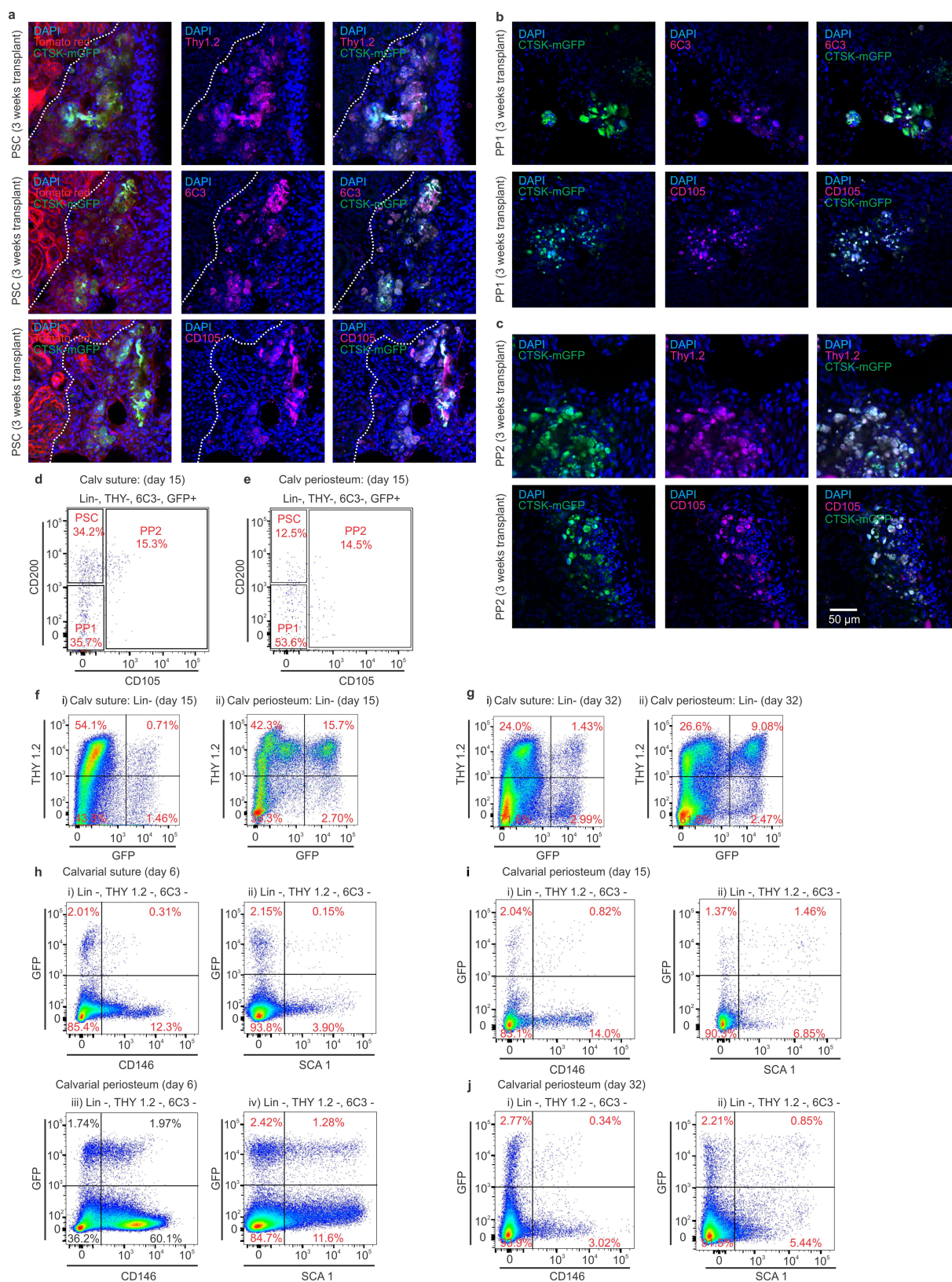
2 weeks after induction of β -actin-Cre with tamoxifen. Enlarged views of outlined regions are shown. Scale bar, 50 μ m. Representative images from 3 independent experiments. **i**, FACS plots showing in vitro differentiation for PP1 (left) and PP2 (right) cells after 15 days of culture. Representative FACS plots from 3 independent experiments. **j**, Alizarin red staining (red) of bone 5 weeks after transplantation of non-CTSK MSCs (left) and PSCs (right) into the kidney capsule. Representative images from 5 independent experiments. Scale bar, 50 μ m. **k**, Relative gene expression for bone- (*Runx2*) and cartilage-specific genes (*Col2a1*, *Comp*, *Chad*) 5 weeks after transplantation of PSCs and non-CTSK MSCs. Non-CTSK MSC-derived cells display significantly higher expression of cartilage specific genes than PSCs. $*P = 0.003$ (*Col2a1*), $*P = 0.002$ (*Comp*), $*P = 0.002$ (*Chad*); two-tailed Student's *t*-test. Data are mean \pm s.d., $n = 3$. **l**, CTSK-mGFP $^{+}$ PSCs (green) were immunostained for Runx2 (magenta, top) and osteocalcin (magenta, middle and bottom) 3, 4 or 5 weeks after transplantation into the kidney capsule. Scale bar, 20 μ m. Representative images from 3 independent experiments.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Functional characterization of non-CTSK MSCs, PSCs and their derivatives. **a**, Total numbers of PSCs and non-CTSK MSCs in mouse femurs at postnatal day 7, day 15 and day 32. Significant decreases in number of PSCs are observed at day 15 (** $P = 0.006$) and day 32 (** $P = 0.009$) compared to day 7. Significant decreases in non-CTSK MSCs are observed at day 15 (*** $P = 3.8 \times 10^{-5}$) and day 32 (*** $P = 0.0003$) compared to day 7. Two-tailed Student's *t*-test. Data are mean \pm s.d., $n = 3$ independent experiments; 5 animals per group for day 7, day 15; 3 animals per group for day 32. **b**, μ CT images of the bone formed by non-CTSK MSCs (left) and PSCs (right) 5 weeks after transplantation. Representative images from 5 independent experiments. **c**, Quantification of bone volume when equal numbers of non-CTSK MSCs and PSCs were transplanted into secondary hosts. Data are mean \pm s.e.m., $n = 3$ independent experiments; two-tailed Student's *t*-test. ns, not significant. **d**, Clonal non-CTSK MSC colonies were split for differentiation into osteoblasts (left, alizarin red staining) and adipocytes (middle, oil red O staining) (scale bar, 20 μ m). Separately, chondrocyte differentiation potential was assayed (right, alcian blue staining; scale bar, 100 μ m). Representative images from 4 independent experiments. **e**, Clonal differentiation capacity of 10 colonies isolated from PSCs and non-CTSK MSCs after subsequent culture under osteogenic (left) and adipogenic (right) differentiation conditions. All 10 colonies from each population were equally multipotent. Data are mean \pm s.d., $n = 3$ independent experiments. **f**, Bright-field images of primary (left),

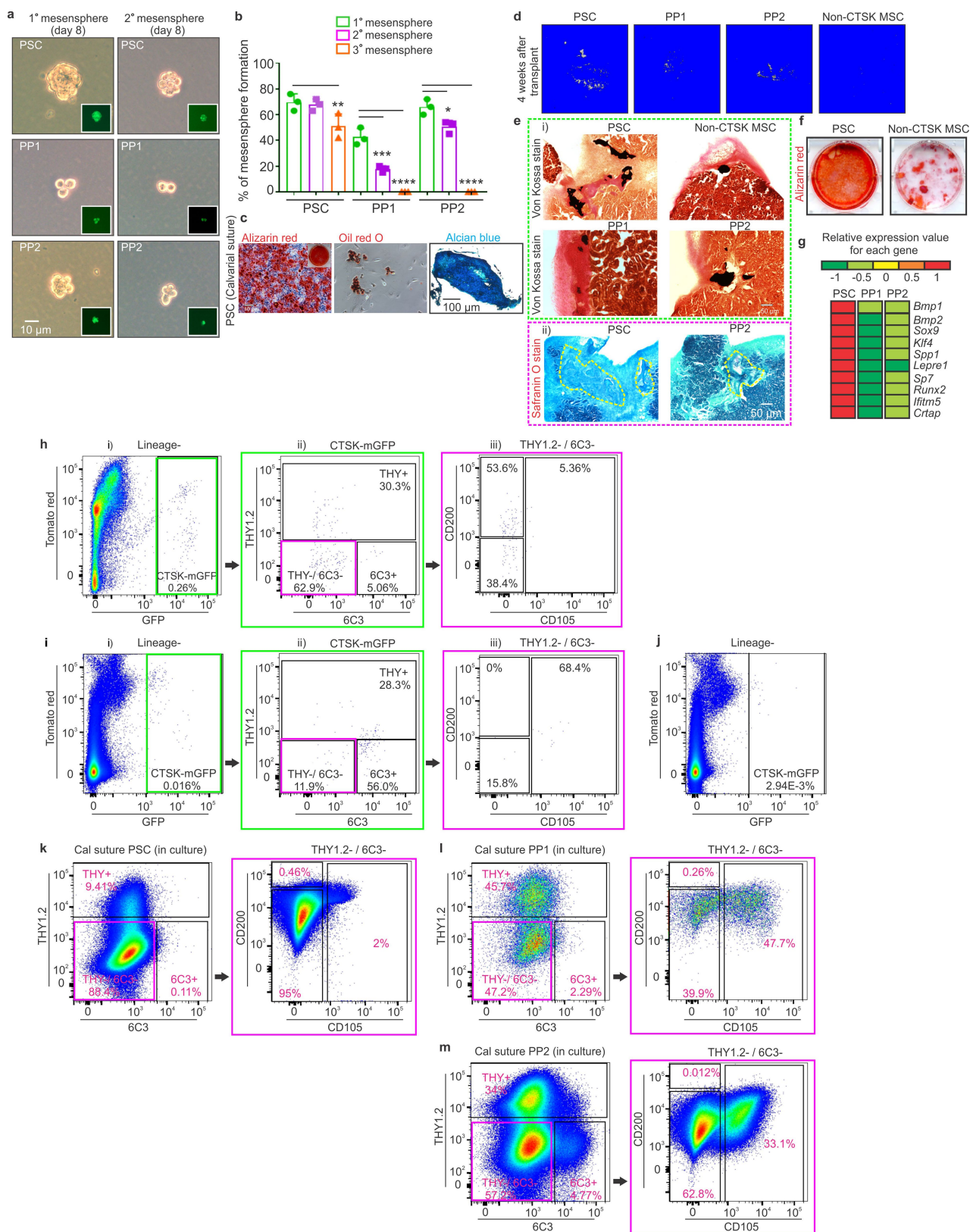
secondary (middle) and tertiary mesenspheres (right) derived from non-CTSK MSCs. Tomato red (red) expression is shown in the insets. Scale bar, 20 μ m. Representative images from 3 independent experiments. **g**, The percentage of PSCs and non-CTSK MSCs able to form mesenspheres. * $P = 0.02$, one-way ANOVA, Dunnett's multiple comparison test. Data are mean \pm s.d., $n = 3$ independent experiments. **h**, FACS analysis of in vitro differentiation of non-CTSK MSCs after 15 days of culture. Red box indicates parent/daughter gate. **i**, FACS plots of non-CTSK MSC-derived cells after the first round of mammary fat pad transplantation. Colour-coded boxes (red and green) indicate parent/daughter gates. FACS plots in **h** and **i** are representative of 3 independent experiments. **j**, FACS for CD140 α (i) and CD146 (ii) in PSCs after transplantation into the mammary fat pad. **k**, FACS for expression of GFP (i), CD140 α (ii), and CD146 (iii) in non-CTSK MSCs after mammary fat pad transplantation. **l**, PP1 cells were transplanted into the mammary fat pad of primary hosts for 2.5 weeks and then analysed by FACS (i–iii). Colour-coded boxes (green and magenta) indicate parent/daughter gates. **m**, **n**, PP2 cells were isolated by FACS and implanted into the mammary fat pad of primary recipients. PP2 derived cells in primary recipients were analysed by FACS (**m**, i–iv), and PP2 cells were re-isolated for transplantation into secondary recipients. PP2-derived cells in secondary recipients were analysed by FACS (**n**, i–iv). Colour-coded boxes (green, magenta and orange) indicate parent/daughter gates. Plots in **j–n** are representative of results from 3 independent experiments.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Differentiation of PSCs, PP1 and PP2 cells when transplanted into kidney capsule of secondary hosts and FACS analysis of CTSK-mGFP calvarial cells. **a**, CTSK-mGFP⁺ PSCs (green) were immunostained for THY1.2 (magenta, top), 6C3 (magenta, middle) and CD105 (magenta, bottom) three weeks after transplantation into the kidney capsule of primary recipients. **b**, CTSK-mGFP⁺ PP1 cells (green) were immunostained for 6C3 (magenta, top) and CD105 (magenta, bottom) three weeks after transplantation into the kidney capsule. **c**, CTSK-mGFP⁺ PP2 cells (green) were immunostained for THY1.2 (magenta, top), and CD105 (magenta, bottom) three weeks after transplantation into the kidney capsule. Scale bar, 50 μ m. Images in **a–c** are representative of 3 independent experiments. Scale bar, 50 μ m.

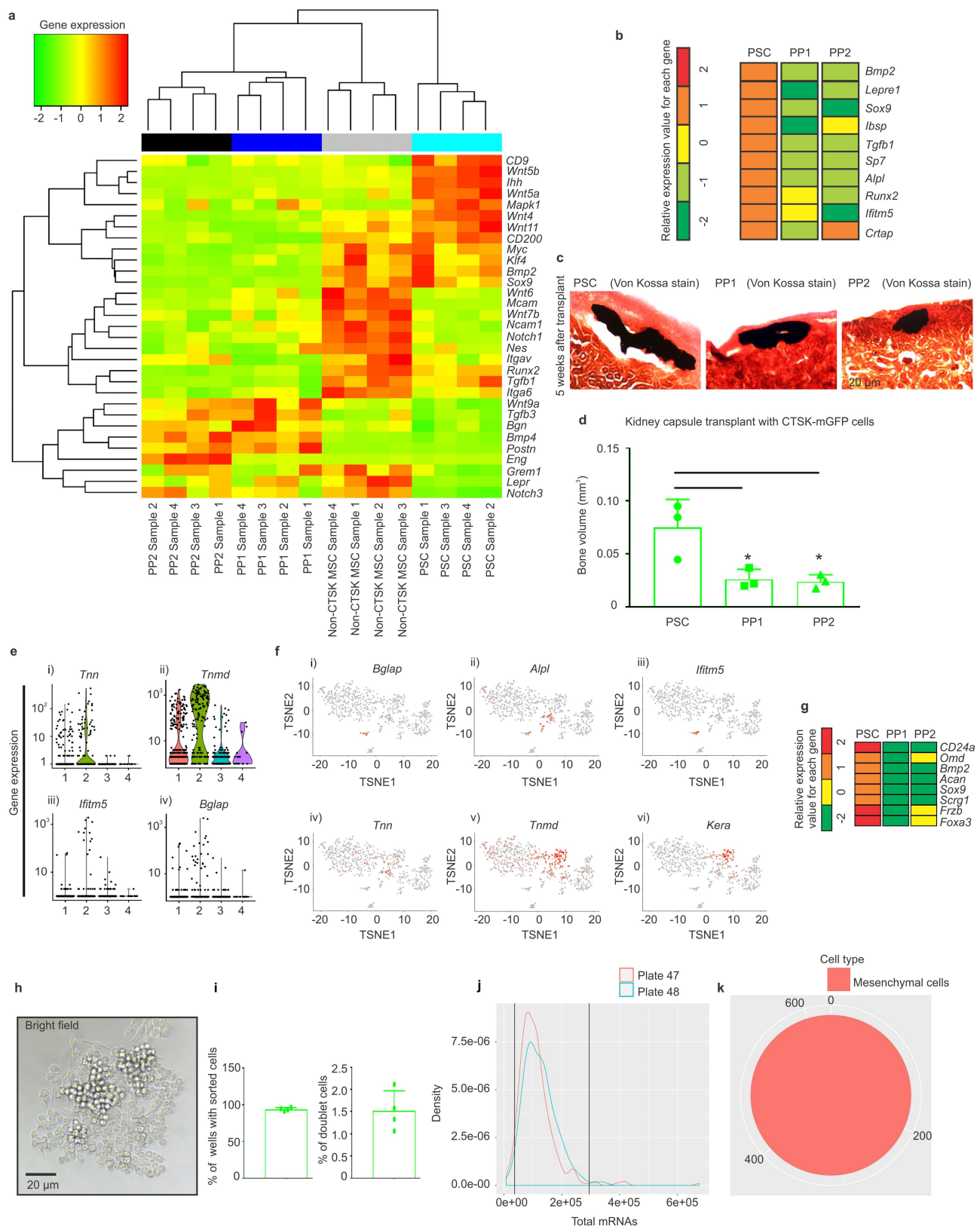
d, e, FACS analysis of PSCs, PP1s and PP2s at P15 in mouse calvarial suture (**d**) and calvarial periosteum (**e**). **f, g**, FACS for THY1.2 in CTSK-mGFP cells isolated from calvarial suture (**f** (i), **g** (i)) and calvarial periosteum (**f** (ii), **g** (ii)) at day 15 (left plots) and day 32 (right plots). Plots in **d–g** are representative of results from 3 independent experiments. **h**, CD146 and SCA1 expression in CTSK-mGFP cells from the suture (i, ii) and periosteum (iii, iv) of P6 mouse calvarium. Representative FACS plots from 10 independent experiments. **i, j**, FACS for CD146 (**i** (i), **j** (i)) and SCA1 (**i** (ii), **j** (ii)) in calvarial periosteum of mice at day 15 (top plots) and day 32 (bottom plots). Representative FACS plots from 3 independent experiments.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Functional characterization of CTSK-mGFP⁺ calvarial suture cells. **a**, Bright-field images of primary (left; scale bar, 10 μ m) and secondary (right) mesenspheres derived from calvarial suture PSCs (top), PP1 (middle) and PP2 (bottom) cells. GFP (green) expression shown in the insets. Representative images from 3 independent experiments. **b**, Quantification of the percentage of PSC, PP1 and PP2 cells able to form mesenspheres. Tertiary colony formation is significantly reduced in PSCs (** $P=0.0034$). PP1 and PP2 cells show significant reduction in both secondary (** $P=0.0002$ for PP1 and * $P=0.016$ for PP2) and tertiary mesensphere formation (**** $P=0.0001$ for PP1 and **** $P=0.0001$ for PP2). One-way ANOVA, Dunnett's multiple comparison test; mean \pm s.d., $n=3$ independent experiments. **c**, Clonal PSC colonies were split for differentiation into osteoblasts (alizarin red staining, left) and adipocytes (oil red O staining, middle; scale bar, 10 μ m). Separately, chondrocyte differentiation potential was assayed (alcian blue staining, right; scale bar, 100 μ m). Representative images from 3 independent experiments. **d**, The amount of bone formed by PSCs, PP1, PP2 and non-CTSK MSCs 4 weeks after transplantation into the kidney capsule of secondary hosts was determined by μ CT. **e**, Von Kossa staining (**e** (i), dotted box in green) displaying bone organoid formation in the

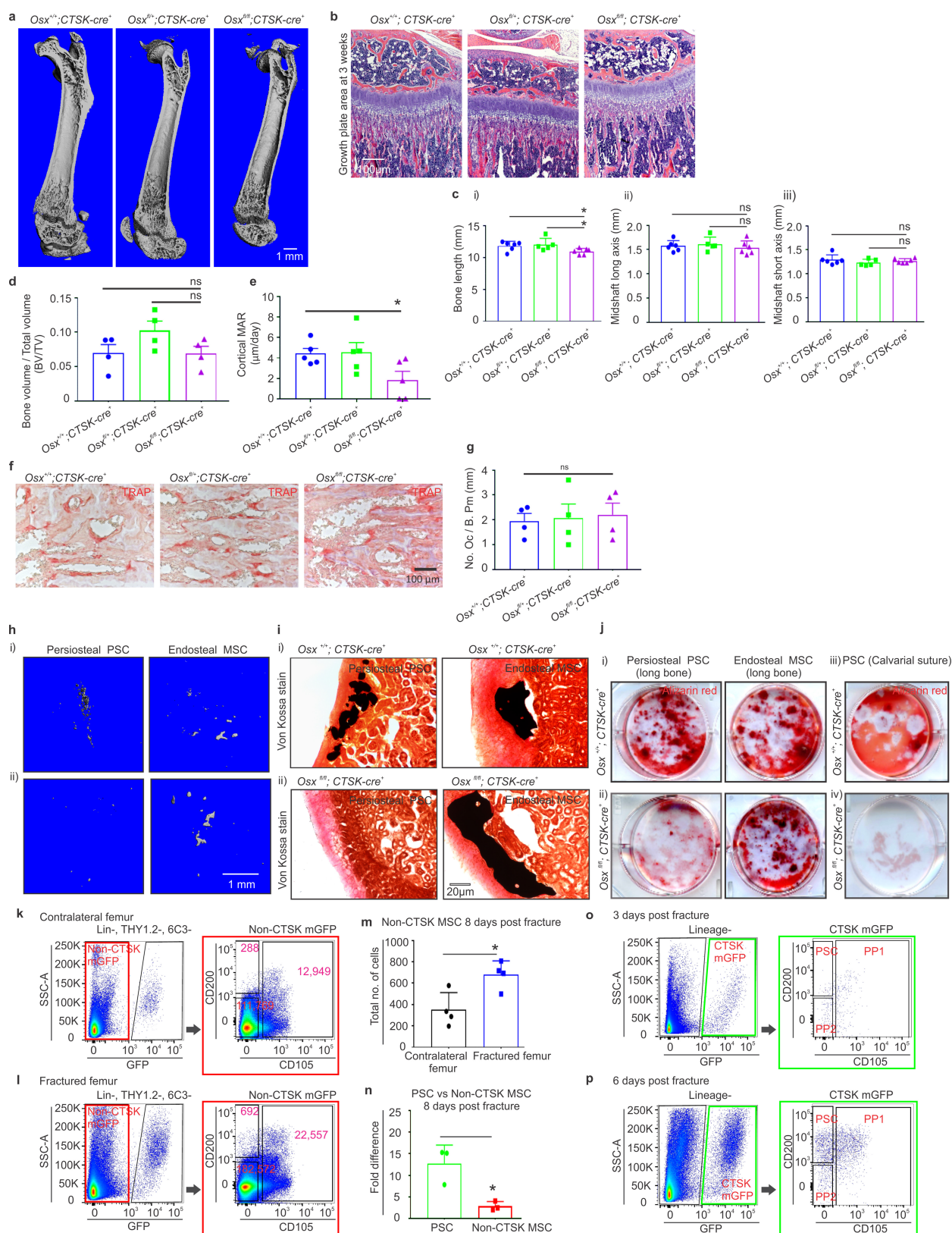
kidney capsule by PSCs (top left), non-CTSK MSCs (top right), PP1 (bottom left) and PP2 (bottom right) cells. Scale bar, 50 μ m. Safranin O staining (**e** (ii), dotted box in magenta) indicates an absence of cartilage formation (transplant area shown by dotted yellow line) after transplant of PSCs (left) and PP2 (right) cells. Scale bar, 50 μ m. **f**, In vitro osteogenic differentiation of PSCs (left) and non-CTSK MSCs (right) as determined by Alizarin red staining (red). Images in **d–f** are representative of 3 independent experiments. **g**, Heat map generated from quantitative real-time PCR analysis shows differences in gene expression between calvarial suture PSCs and the progenitor populations, PP1 and PP2 cells. **h**, PSCs were transplanted into a mammary fat pad of primary hosts for 2.5 weeks and then analysed by FACS (i–iii). Colour-coded boxes (green and magenta) indicate parent/daughter gates. **i**, FACS analysis of PP2 cells after transplantation into the mammary fat pad of primary hosts (i–iii). **j**, FACS analysis shows poor engraftment and loss of PP1 cells (as detected by GFP expression) when transplanted into the mammary fat pad of primary hosts. **k–m**, FACS plots demonstrating the in vitro differentiation of PSC (**k**), PP1 (**l**), and PP2 cells (**m**) when cultured for 2 weeks. Magenta boxes indicate parent/daughter gates for each analysed cell population. Plots in **h–m** represent results from 3 independent experiments.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Gene expression analysis in CTSK-mGFP cells isolated from mouse femurs. **a**, Bulk RNA-seq analysis of FACS-isolated PSC, PP1, PP2 and non-CTSK-mGFP MSCs from 6-day-old mouse femurs. Hierarchical clustering analysis was performed on RNA-seq data. **b**, Heat map generated from bulk RNA-seq of FACS-sorted cells shows differences in gene expression between PSCs and the progenitor populations, PP1 and PP2 cells. **c**, Von Kossa staining indicates bone organoid formation by PSCs (left), PP1 (middle) and PP2 cells (right) 5 weeks after transplantation into the kidney capsule. Scale bar, 20 μ m. Representative images from 3 independent experiments with 3 mice per group. **d**, Significantly reduced bone formation (bone volume) in PP1 (* $P=0.04$) and PP2 (* $P=0.032$) cells compared to PSCs after transplantation. Two-tailed Student's t -test. Data are mean \pm s.d., $n=3$ independent experiments. **e**, Relative expression of *Tnn* (i), *Tnmd* (ii), *Ifitm5* (iii) and *Bglap* (iv) among the four clusters (identified by 1–4) that

were generated through analysis of 658 CTSK-mGFP⁺ mesenchymal cells using the Seurat package. Cell clusters (1–4) along the x axis. **f**, Expression of genes such as *Bglap* (i), *Alpl* (ii), *Ifitm5* (iii), *Tnn* (iv), *Tnmd* (v) and *Kera* (vi) are shown by pseudocolouring of t -SNE plots. **g**, Heat map generated from bulk RNA-seq shows differences in gene expression between PSCs and the progenitor populations, PP1 and PP2 cells. **h–k**, Monocle analysis of CEL-Seq2 data. **h**, Bright-field image of a colony that was generated from single-cell sorting of RAW264.7 cells by FACS. Scale bar, 20 μ m. Representative image from 3 independent experiments. **i**, Graphs indicate the percentage of wells that received sorted cells by FACS (left) and the percentage of doublets detected in those wells (right). Data are mean \pm s.d., $n=4$. **j**, Data represent the total amount of mRNA in the two 384-well plates (plate 47 and plate 48) that were sequenced using CEL-Seq2. **k**, Pie chart, showing that the analysed CTSK-mGFP⁺ cells were mesenchymal in origin.

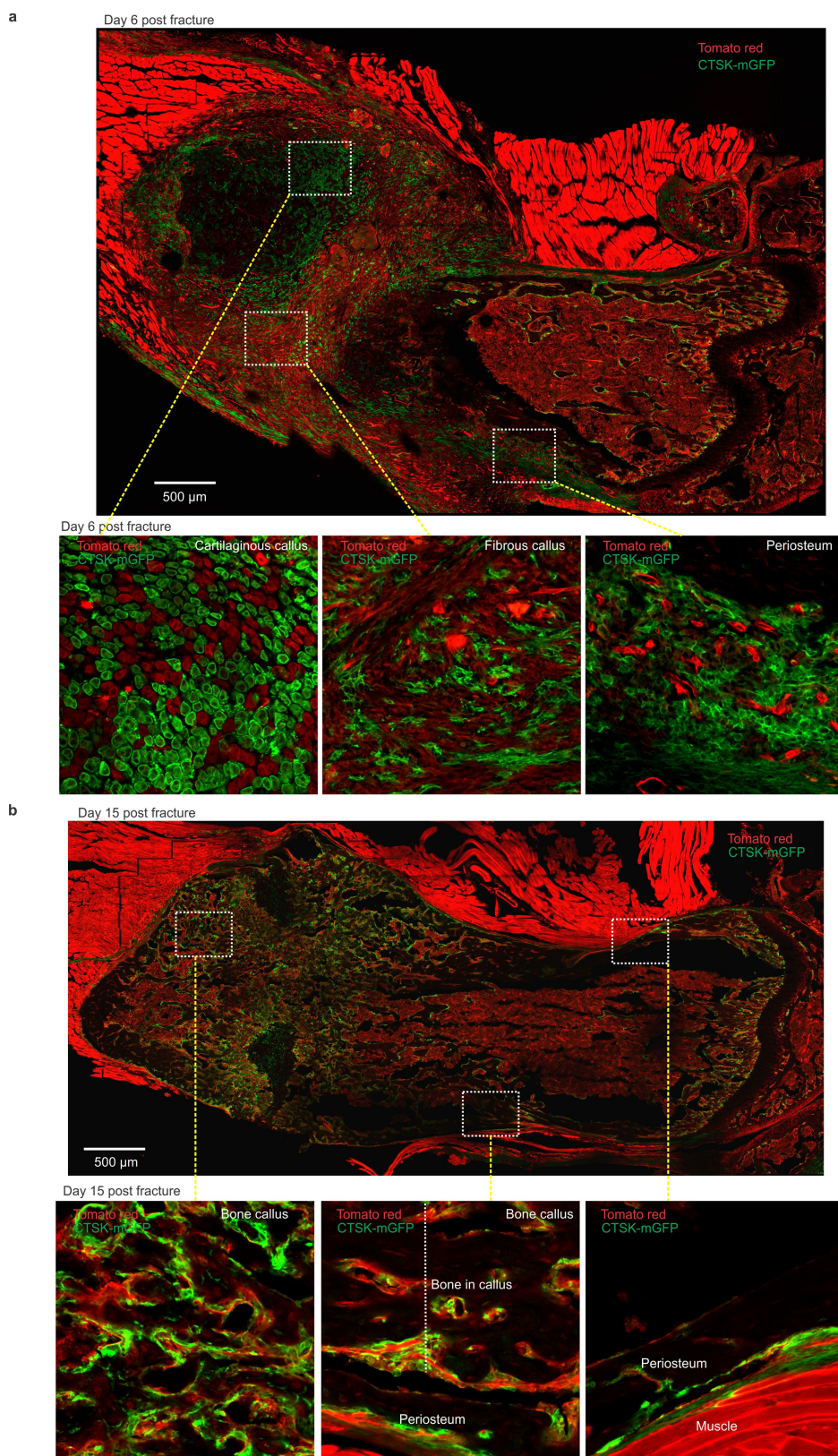


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | μ CT, histomorphometric analysis and characterization of cells isolated from $Osx^{fl/fl};Ctsk^{cre}$ mouse femur.

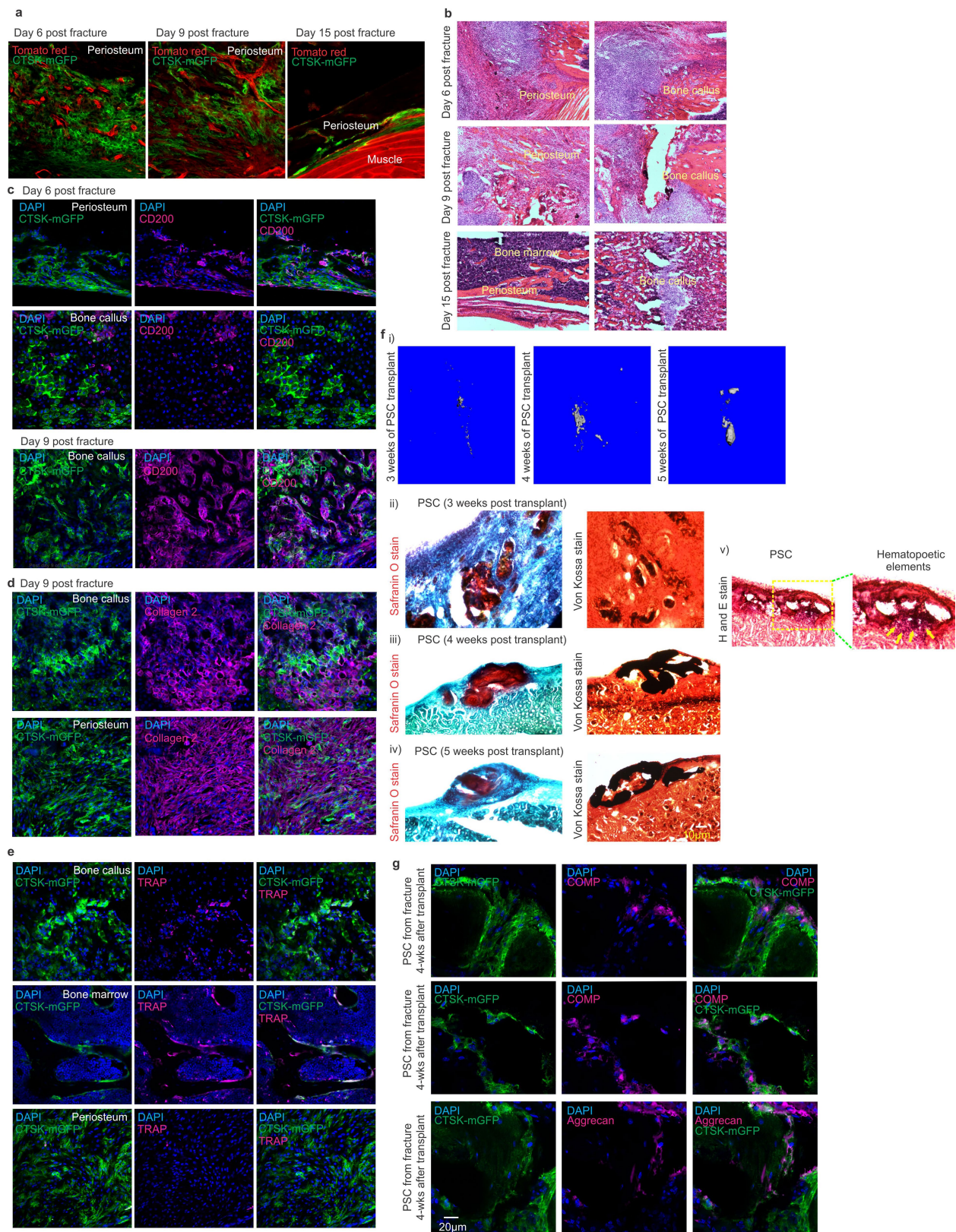
a, μ CT images of longitudinal sections of femurs from $Osx^{fl/fl};Ctsk^{cre}$ mice or littermate controls at 4 weeks of age. **b**, Haematoxylin and eosin staining showing growth plate morphology in $Osx^{fl/fl};Ctsk^{cre}$ mice or littermate controls. Images in **a** and **b** are representative of 5 independent experiments. Scale bar, 100 μ m. **c**, Bone length (i), midshaft along long axis (ii) and midshaft along short axis (iii). $Osx^{fl/fl};Ctsk^{cre}$ mice show a significant reduction in bone length (i) compared to $Osx^{fl/+};Ctsk^{cre}$ (* $P=0.039$) and $Osx^{+/+};Ctsk^{cre}$ (* $P=0.034$). Two-tailed Student's t -test. Data are mean \pm s.d., $n=6$ animals per group. **d**, Bone volume/total volume (BV/TV) for trabecular bone. Data are mean \pm s.e.m., $n=4$ animals per group, two-tailed Student's t -test. **e**, Histomorphometric parameters. Cortical mineral apposition rate (MAR; μ m day $^{-1}$). * $P=0.031$; two-tailed Student's t -test; data are mean \pm s.e.m., $n=5$ animals per group at 4 weeks of age. **f**, TRAP staining of osteoclasts in the trabecular bone area of femurs of the indicated mice at 4 weeks of age. Scale bar, 100 μ m. Representative images from 4 independent experiments. **g**, Quantification of osteoclast number/bone perimeter (No. Oc/B. Pm). Data are mean \pm s.e.m., $n=4$ animals per group, two-tailed Student's t -test. **h**, μ CT images showing the amount of bone formed when periosteal PSCs (left column) and endosteal MSCs (right column) isolated from femurs of $Osx^{+/+};Ctsk^{cre}$ (top) and $Osx^{fl/fl};Ctsk^{cre}$ mice (bottom) were transplanted into the kidney capsule. Scale bar, 1 mm. **i**, Von Kossa staining (black) of bone organoids formed after

transplantation of periosteal PSCs (left column) and endosteal MSCs (right column) isolated from $Osx^{+/+};Ctsk^{cre}$ (top) and $Osx^{fl/fl};Ctsk^{cre}$ mice (bottom) and transplanted into the kidney capsule. Scale bar, 20 μ m. Images in **h** and **i** are representative of 3 independent experiments. **j**, Alizarin red staining (red) of periosteal PSCs (left column) and endosteal MSCs (right column) isolated from the femur (i, ii) and calvarial sutures (iii, iv) of $Osx^{+/+};Ctsk^{cre}$ (top panel) and $Osx^{fl/fl};Ctsk^{cre}$ mice (bottom panel) after culture under osteoblast differentiation conditions. Images are representative of 3 independent experiments. **k**, **l**, FACS plots of contralateral unfractured femurs (**k**) and fractured femurs (**l**). Colour-coded boxes (red) indicate parent/daughter gates. Representative FACS plots from 3 independent experiments. **m**, A significant increase (* $P=0.019$) is seen in non-CTSK MSCs in callus tissue 8 days post fracture. Values displayed represent the absolute number of cells isolated per fracture callus. Data are mean \pm s.d., $n=4$ independent experiments, 4 animals/group; two-tailed Student's t -test. **n**, Graph displays significantly (* $P=0.017$) higher fold PSC count than non-CTSK MSCs in the fractured callus. Data are mean \pm s.d.; $n=3$ independent experiment; two-tailed Student's t -test. Values displayed are normalized relative to the pre-fracture numbers of the same corresponding population to demonstrate the fold expansion after fracture. **o**, **p**, FACS of cells from fractured callus after 3 (**o**) and 6 days (**p**) of fracture. Colour-coded boxes (green) indicate parent/daughter gates. Representative FACS plots from 3 independent experiments.



Extended Data Fig. 8 | *Ctsk^{cre};mTmG* mouse femur 6 days and 15 days after fracture. a, Mouse femur 6 days after fracture (top). Bottom, enlarged view of areas indicated by dotted white boxes. CTSK-mGFP (green), mTomato red (red). Scale bar, 500 μm. Images are representative

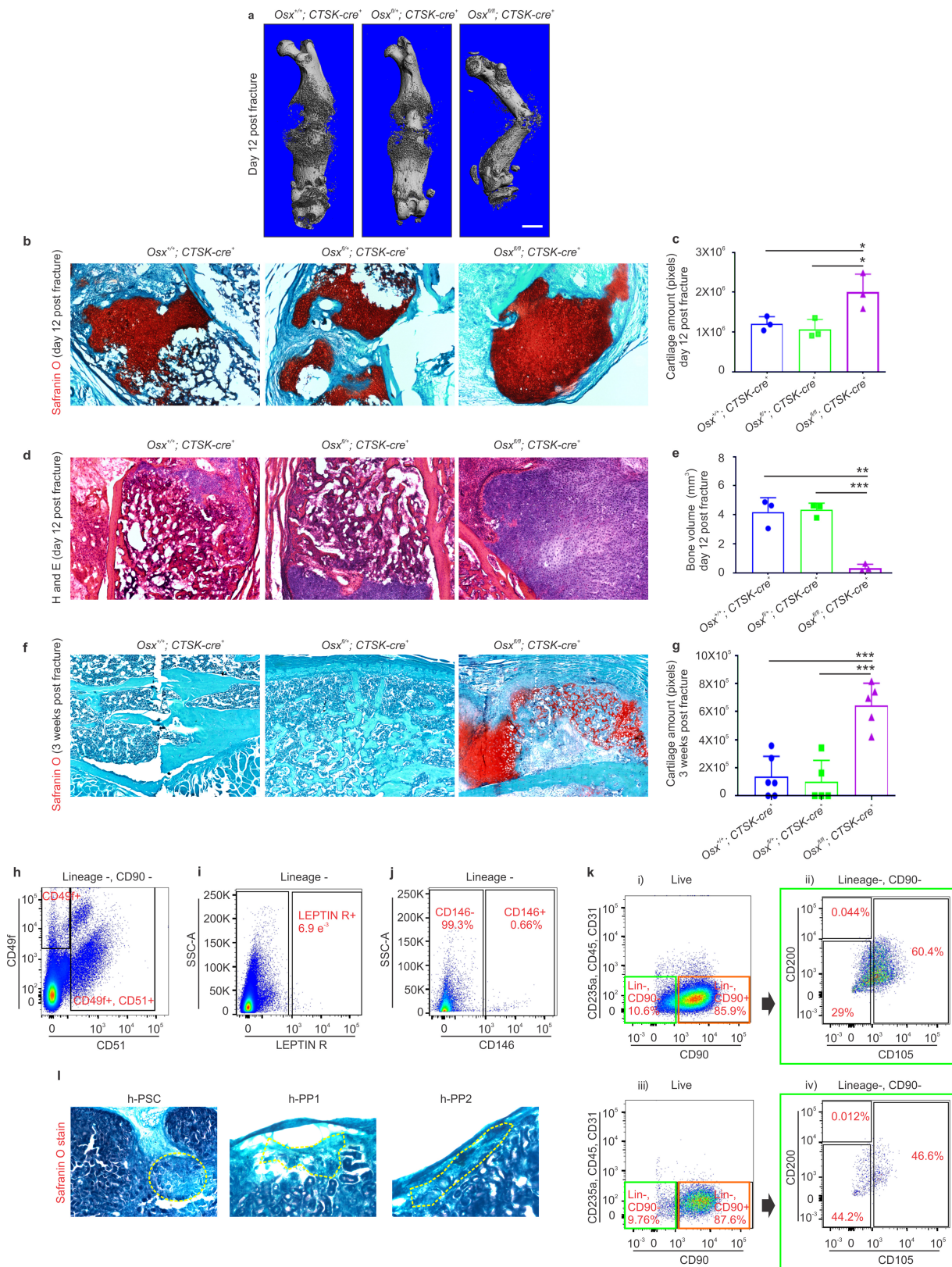
of 3 independent experiments. b, Mouse femur 15 days after fracture (top). Bottom, enlarged view of areas indicated by dotted white boxes. Scale bar, 500 μm. Images are representative of 3 independent experiments.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Characterization of CTSK-mGFP cells of mouse femur after fracture. **a**, The periosteum of mouse femur 6 (left), 9 (middle) and 15 (right) days after fracture. **b**, Haematoxylin and eosin staining of callus tissue 6 (top), 9 (middle) and 15 (bottom) days after fracture. **c**, CD200 (magenta) immunostaining of femurs collected 6 (top and middle) and 9 (bottom) days after fracture. **d**, Immunostaining for type 2 collagen (magenta) 9 days after fracture. **e**, TRAP staining (magenta), identifying osteoclasts in the bone callus (top) and bone marrow (middle) of fractured femurs. Few to no TRAP-positive cells were present in the periosteal region (bottom). Images in **a–e** are representative of 3 independent experiments. **f**, PSCs isolated from fracture callus were transplanted into kidney capsule secondary hosts. μ CT images

of bone formation at 3 (left), 4 (middle) and 5 weeks (right) after PSC transplantation to the kidney capsule (i). Safranin O staining (red), and Von-Kossa staining (black) were performed on sectioned kidney samples to detect cartilage and bone at 3 (ii), 4 (iii) and 5 (iv) weeks after PSC transplantation. Scale bar, 10 μ m. Haematoxylin and eosin staining indicates that PSCs isolated from fracture callus are competent to recruit host-derived haematopoietic elements at the site of transplantation (yellow arrows, v). **g**, Immunostaining reveals co-localization of CTSK-mGFP⁺ cells (green) with cartilage-specific markers such as COMP (magenta, top and middle) and aggrecan (magenta, bottom) 4 weeks after PSC transplantation. Scale bar, 20 μ m. Images in **f** and **g** are representative of 3 independent experiments.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Characterization of *Osx^{fl/fl};Ctsk^{cre}* mice and human periosteal cells. **a**, μ CT images of *Osx^{fl/fl};Ctsk^{cre}* mice 12 days post-fracture. Scale bar 1 mm. **b**, Safranin O staining was performed to detect cartilage in the callus 12 days after fracture. Images in **a** and **b** are representative of 3 independent experiments. **c**, Significantly higher amounts of cartilage were detected in *Osx^{fl/fl};Ctsk^{cre}* mice compared to *Osx^{fl/+};Ctsk^{cre}* (* $P=0.035$) and *Osx^{+/+};Ctsk^{cre}* (* $P=0.04$) mice. Two-tailed Student's t -test; data are mean \pm s.d., $n=3$ independent experiments, 3 mice per group. **d**, Haematoxylin and eosin staining of callus tissue 12 days after fracture. Representative images from 3 independent experiments. **e**, Significantly lower bone volume (BV) was detected in *Osx^{fl/fl};Ctsk^{cre}* mice compared to *Osx^{fl/+};Ctsk^{cre}* (** $P=0.0002$) and *Osx^{+/+};Ctsk^{cre}* (** $P=0.002$) mice. Two-tailed Student's t -test; data are mean \pm s.d., $n=3$ independent experiments, 3 animals per group. **f**, Safranin O staining for callus tissue 3 weeks after fracture. Representative images from 3 independent experiments. **g**, Significantly higher amounts of

cartilage were detected in *Osx^{fl/fl};Ctsk^{cre}* mice compared to *Osx^{fl/+};Ctsk^{cre}* (** $P=0.0005$) and *Osx^{+/+};Ctsk^{cre}* (** $P=0.0003$) mice at 3 weeks after fracture. Two-tailed Student's t -test; data are mean \pm s.d., $n=3$ independent experiments; 6 mice for control, 5 mice each for heterozygote and knockout. **h–i**, FACS using CD49f, CD51 (**h**) LEPR (**i**) and CD146 (**j**) in human periosteal cells obtained from the femur. Representative FACS plots from 10 independent experiments. **k**, In vitro differentiation of h-PSCs (**k**, i, ii) and h-PP1 cells (**k**, iii, iv) after 3 weeks of culture. Colour-coded boxes (green) indicate parent/daughter gates for each cell type. Representative FACS plots from 3 independent experiments. **l**, Safranin O staining showing an absence of cartilage formation after h-PSC (left), h-PP1 (middle) and h-PP2 (right) transplantation into the kidney capsule of immunocompromised mice. The area containing the transplanted tissue is shown by the dotted yellow line. Representative images from 3 independent experiments.

A flavin-based extracellular electron transfer mechanism in diverse Gram-positive bacteria

Samuel H. Light¹, Lin Su^{2,3}, Rafael Rivera-Lugo¹, Jose A. Cornejo², Alexander Louie¹, Anthony T. Iavarone⁴, Caroline M. Ajo-Franklin² & Daniel A. Portnoy^{1,5*}

Extracellular electron transfer (EET) describes microbial bioelectrochemical processes in which electrons are transferred from the cytosol to the exterior of the cell¹. Mineral-respiring bacteria use elaborate haem-based electron transfer mechanisms^{2–4} but the existence and mechanistic basis of other EETs remain largely unknown. Here we show that the food-borne pathogen *Listeria monocytogenes* uses a distinctive flavin-based EET mechanism to deliver electrons to iron or an electrode. By performing a forward genetic screen to identify *L. monocytogenes* mutants with diminished extracellular ferric iron reductase activity, we identified an eight-gene locus that is responsible for EET. This locus encodes a specialized NADH dehydrogenase that segregates EET from aerobic respiration by channelling electrons to a discrete membrane-localized quinone pool. Other proteins facilitate the assembly of an abundant extracellular flavoprotein that, in conjunction with free-molecule flavin shuttles, mediates electron transfer to extracellular acceptors. This system thus establishes a simple electron conduit that is compatible with the single-membrane structure of the Gram-positive cell. Activation of EET supports growth on non-fermentable carbon sources, and an EET mutant exhibited a competitive defect within the mouse gastrointestinal tract. Orthologues of the genes responsible for EET are present in hundreds of species across the Firmicutes phylum, including multiple pathogens and commensal members of the intestinal microbiota, and correlate with EET activity in assayed strains. These findings suggest a greater prevalence of EET-based growth capabilities and establish a previously underappreciated relevance for electrogenic bacteria across diverse environments, including host-associated microbial communities and infectious disease.

L. monocytogenes is a fermentative Gram-positive bacterium that is frequently associated with decaying plant matter in the environment, but which transforms into an intracellular pathogen on encountering a mammalian host⁵. Despite lacking a lifecycle or genes conventionally associated with EET, a 25-year-old observation that *L. monocytogenes* possessed extracellular ferric iron reductase activity⁶ led us to wonder whether a novel EET strategy existed. Because electrons transferred out of the cell can be captured by an electrode, electrochemical measurements provide a useful tool for assaying EET⁷. By performing chronoamperometry experiments, we observed that *L. monocytogenes* produces a robust electric current in the presence of growth substrate (Fig. 1a, Extended Data Fig. 1a). In addition, we found that cyclic voltammetry experiments—which monitor electric current while the electrochemical potential is systematically varied—revealed a distinctive catalytic wave reminiscent of other electrochemically active bacteria^{8,9} (Extended Data Fig. 1b). These results thus provide strong evidence that *L. monocytogenes* possesses EET activity.

To address the genetic basis of EET activity, approximately 50,000 colonies of a pooled *L. monocytogenes* *himar1* transposon library were grown on Fe³⁺-containing agar plates. Mutants with decreased

colorimetric change following an Fe²⁺-indicator overlay were visually identified and the location of their transposon insertion was mapped to the genome (Fig. 1b). From this screen, thirty-four independent transposon insertions that localized to a largely uncharacterized 8.5-kilobase locus were identified—with at least one insertion disrupting each of the eight genes in this region (Fig. 1c). Genes in the locus were assigned names on the basis of putative functions of their protein products (see ‘Gene name assignment’ in Methods for a more detailed explanation). The only transposon insertions outside the identified locus disrupt *ribU*, the substrate-binding subunit of a riboflavin transporter¹⁰

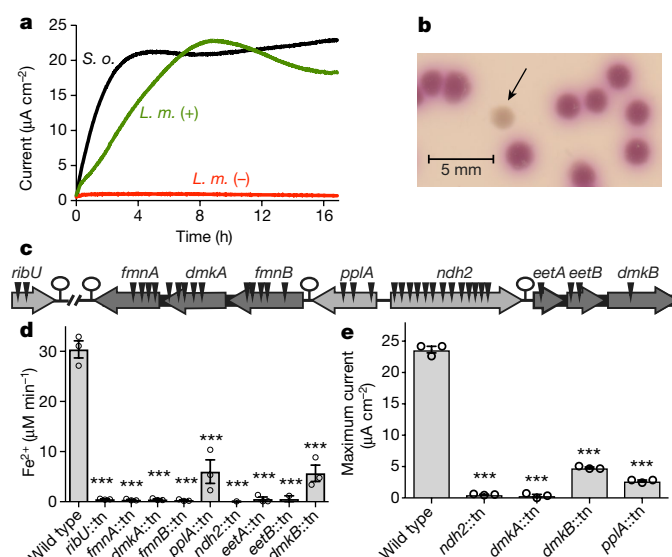


Fig. 1 | An uncharacterized genetic locus associated with EET activity. **a**, Chronoamperometry results from *L. monocytogenes* (*L. m.*)- or *Shewanella oneidensis* (*S. o.*)-inoculated electrochemical reactors. For *L. monocytogenes* experiments, an electron donor (glucose) was present in (+) or absent from (–) the medium; lactate was used as an electron donor for *S. oneidensis*. Results are representative of three independent experiments. **b**, Image of one of the thirty-six independent mutants identified from the ferric iron reduction screen, minutes after ferrozine agar overlay (arrow). **c**, Location of transposon insertions (triangles) in mutants identified as having decreased ferric iron reductase activity in the genetic screen. Arrows represent genes, with the darker grey signifying inclusion in a multi-gene operon. Previously uncharacterized genes on the locus have been assigned names based on the putative functions of the proteins they encode. **d**, Ferric iron reductase activity of transposon mutants identified from the screen. Results are expressed as mean \pm s.e.m. from three independent experiments. **e**, Maximum electric current achieved from chronoamperometry experiments with representative EET mutants. Strains that statistically differ from the wild-type strain are indicated; *** $P < 0.001$, ANOVA with Dunnett's post-test.

¹Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. ²Molecular Foundry, Molecular Biophysics and Integrated Bioimaging, and Synthetic Biology Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ³State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210018, China. ⁴QB3/Chemistry Mass Spectrometry Facility, University of California, Berkeley, Berkeley, CA, USA. ⁵Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA. *e-mail: portnoy@berkeley.edu

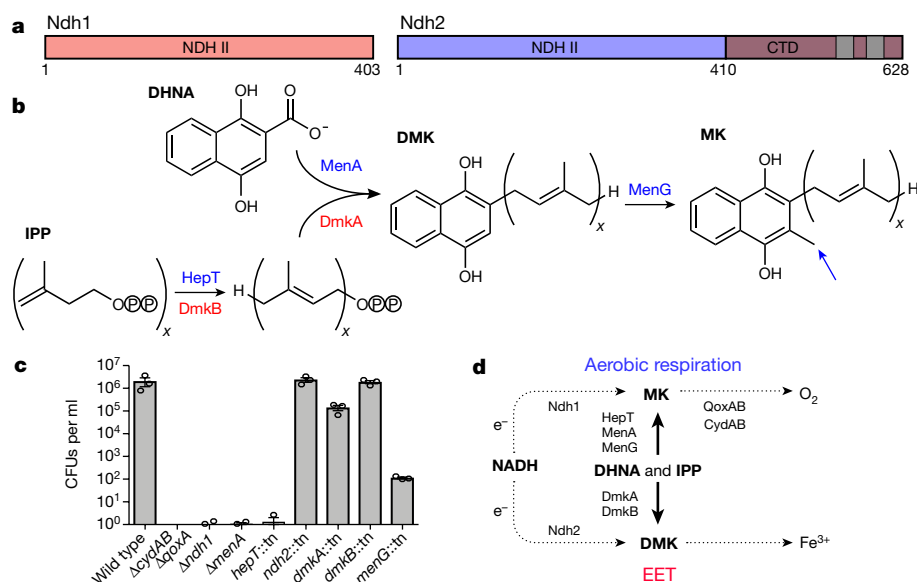


Fig. 2 | A parallel electron transfer pathway segregates EET from aerobic respiration. **a**, Domain layout of the *L. monocytogenes* proteins Ndh1 and Ndh2. CTD, C-terminal domain; NDH II, type II NADH dehydrogenase domain. Grey regions represent predicted transmembrane helices. **b**, Predicted reactions catalysed by *L. monocytogenes* DmkA and DmkB, the paralogous proteins MenA and HepT, and MenG (highlighted by blue arrow). DHNA, 1,4-dihydroxy-2-naphthoyl-CoA; DMK, demethylmenaquinone; IPP, isopentenyl pyrophosphate; MK, menaquinone; x, an unknown number of isoprene repeats (which may

differ between the two quinones). **c**, Colony-forming units (CFUs) after 24 h in aerobic respiration medium. Results from three independent experiments are expressed as mean \pm s.e.m. The Δ cydAB Δ qoxA mutant lacks terminal cytochrome oxidases and thus provides an aerobic-respiration-deficient control. **d**, Probable electron transfer pathways inferred from mutants with EET (red) or aerobic respiration (blue) phenotypes. Dashed arrows highlight the path of electron flow and solid lines track quinone synthesis.

(Fig. 1c). We confirmed that the mutants had diminished ferric iron reductase (Fig. 1d) and electrochemical activity (Fig. 1e, Extended Data Fig. 1b) and then turned to study the molecular basis of EET.

Type II NADH dehydrogenase—or Ndh1 in *L. monocytogenes*—catalyses electron exchange from cytosolic NADH to a lipid-soluble quinone derivative, which is the first step in the respiratory electron transport chain¹¹. Ndh2, which is encoded by one of the genes in the EET locus, is a protein with an N-terminal type II NADH dehydrogenase domain and a unique transmembrane C-terminal domain that is absent from functionally characterized enzymes (Fig. 2a). Consistent with Ndh2 being a novel NADH dehydrogenase, we observed that EET activation correlated with cellular NAD⁺ levels (Extended Data Fig. 2). Furthermore, the proteins DmkA and DmkB—which are encoded by two other genes in the EET locus—are paralogues of the highly conserved microbial enzymes MenA and HepT, which catalyse terminal steps in the production of the quinone demethylmenaquinone (Fig. 2b). In *Escherichia coli*, three different quinones—demethylmenaquinone, menaquinone and ubiquinone—are used to selectively channel electrons to different electron acceptors¹². By analogy, we reasoned that a distinct quinone derivative and NADH dehydrogenase might functionally segregate electron fluxes for EET and aerobic respiration.

To clarify the relationship between EET and aerobic respiration, we formulated an ‘aerobic respiration medium’ that contained non-fermentable glycerol as the sole carbon source. Despite exhibiting wild-type levels of ferric iron reductase activity (Extended Data Fig. 3a), Δ cydAB Δ qoxA (a positive control that lacks terminal cytochrome oxidases), Δ menA, Δ hepT::tn and Δ ndh1 strains failed to grow on aerobic respiration medium (Fig. 2c). By contrast, EET mutants grew similarly to wild-type strains under these conditions (Fig. 2c). Moreover, Δ menG—which encodes the enzyme that converts demethylmenaquinone to menaquinone—is contained on an operon with Δ hepT and is essential for growth on aerobic respiration medium, but not ferric iron reductase activity (Fig. 2c, Extended Data Fig. 3). Collectively, these results support the conclusion that a demethylmenaquinone derivative used by Ndh2 and a menaquinone derivative used by Ndh1 are selective for downstream enzymes that function in EET and aerobic respiration, respectively (Fig. 2d).

We next sought to address the downstream steps responsible for electron transfer from the quinone pool to extracellular electron acceptors. FmnB is a predicted lipoprotein that is annotated as possessing FMN transferase activity. Homologous FMN transferases catalyse a post-translational modification in which an FMN moiety is covalently linked to a threonine side chain of substrate proteins^{13,14} (Fig. 3a). To identify protein substrates of FmnB, wild-type and Δ fmnB::tn cells were subjected to a comparative mass spectrometric analysis. Only two *L. monocytogenes* peptides met the criteria of selective FMNylation in the wild-type sample and both of these mapped to distinct regions in the protein product of the neighbouring gene in the EET locus, PplA (Supplementary Table 1).

Similar to FmnB, PplA is a predicted lipoprotein and, consistent with this prediction, a trypsin-shaving experimental approach, in which extracellular-surface-associated proteins liberated through a partial digestion of the cell wall are identified by mass spectrometry, confirmed that PplA is associated with the surface of the cell (Supplementary Table 2). The N-terminal lipidation site on PplA is followed by approximately 30 amino acids that are predicted to be unstructured. N-terminal unstructured regions are a common feature of bacterial lipoproteins and are thought to provide a loose tether that allows the active portion of the protein to diffuse further from the membrane and to partially or fully penetrate the cell wall¹⁵. This property, coupled with the covalently bound redox-active FMNs, is consistent with PplA representing the extracellular component of the EET machinery that facilitates electron transfer—via its FMNs—to extracellular electron acceptors.

Following its unstructured N-terminal region, PplA has sequential domains that share 59% sequence identity with each other. From the proteomic analysis, it is evident that the FMNylated threonines on PplA assume equivalent positions on each of these related domains (Fig. 3b). To further clarify the mechanism of FMNylation, we tested FmnB substrate specificity using recombinant FmnB and PplA. These assays confirm that FmnB catalyses FMNylation of PplA and demonstrate that the enzyme specifically uses flavin adenine dinucleotide (FAD) as a substrate (Fig. 3c, Extended Data Fig. 4).

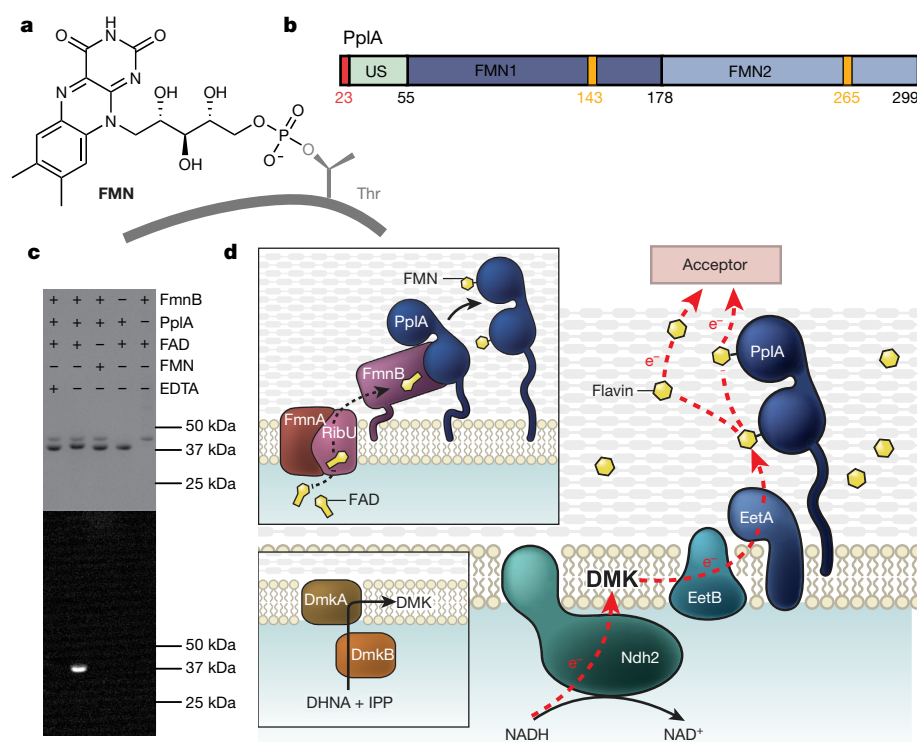


Fig. 3 | A surface-associated flavoprotein establishes the extracellular component of the EET apparatus. a, Post-translational modification catalysed by the FMN transferase family of enzymes, of which FmnB is a member^{13,14}. **b**, Domain architecture of PplA. FMN1, FMNylated domain 1; FMN2, FMNylated domain 2; US, unstructured. The lipidated cysteine on the N terminus after signal peptidase processing is shown in red and FMNylated threonines are shown in yellow. **c**, Analysis of FmnB substrate specificity. SDS-PAGE of recombinant PplA after incubation under specified conditions. Ultraviolet illumination of the gel (bottom) enables visualization of protein with covalently bound flavin. Results are

representative of three independent experiments. See Supplementary Fig. 1 for uncropped gel. **d**, Model of the molecular basis of EET. DmkA and DmkB synthesize a demethylmenaquinone derivative (bottom inset). RibU and FmnA secrete FAD that is used by FmnB to post-translationally modify PplA (top inset). EET is achieved by a series of electron transfers. Ndh2 transfers electrons from NAD to DMK. Electrons are transferred from DMK to FMN groups on PplA or free flavin shuttles—possibly with involvement from uncharacterized membrane proteins in the EET locus, EetA and EetB—and ultimately to a terminal electron acceptor.

Because both FmnB and PplA are membrane-anchored lipoproteins, a source of FAD substrate is required for FmnB to FMNylate PplA on the surface of the cell. The only transposon insertions identified outside the EET locus disrupt *ribU*, which has previously been shown to encode the substrate-binding subunit of an energy-coupling factor (ECF) transporter that functions in riboflavin uptake¹⁰. In addition to a substrate-binding subunit, ECF transporters contain a transmembrane subunit and two distinct ATPase subunits, which drive the transport of substrate across the membrane¹⁰ (Extended Data Fig. 5a). FmnA in the EET locus shares 50% sequence identity with the transmembrane subunit of the RibU–ECF riboflavin transporter (EcT) and this led us to propose that FmnA interacted with RibU to promote FAD secretion (Extended Data Fig. 5b). Consistent with this interpretation, proteomic analysis of *ribU::tn* and *fmnA::tn* strains revealed a marked decrease in PplA FMNylation (Supplementary Table 1). Furthermore, addition of FAD to the growth medium specifically restored ferric iron reductase activity to the *ribU::tn* and *fmnA::tn* strains (Extended Data Fig. 5c). On the basis of these findings, we propose that RibU and FmnA establish a transporter that secretes the FAD required for FmnB-catalysed FMNylation of PplA.

The term ‘extracellular electron shuttle’ refers to redox-active small molecules that are cyclically reduced by cells and oxidized by extracellular electron acceptors^{16,17}. The relevance of shuttles for EET is exemplified by *Shewanella* species, which use an efflux-type transporter to secrete flavins that shuttle electrons to acceptors that are not directly contacting the cell^{18–20}. In contrast to *Shewanella*, *L. monocytogenes* is a flavin auxotroph and thus by definition environmental flavins must be present in its replicative niche. Indeed, micromolar flavin concentrations are typical of nutrient-rich environments, such as the plant biomass and mammalian hosts in which

L. monocytogenes proliferates^{21,22}. To determine whether flavins could be used as electron shuttles, we tested the effect of exogenous riboflavin, FMN and FAD on EET activity. Injection of FMN into an *L. monocytogenes*-inoculated electrochemical chamber resulted in a pronounced increase in electric current (Extended Data Fig. 6a). Moreover, while cells immersed in soluble ferric iron exhibited a high baseline level of reductase activity that was unresponsive to flavins, flavins caused a marked concentration-dependent enhancement in the reduction of insoluble ferric (hydr)oxide (Extended Data Fig. 6b). These data thus support the conclusion that *L. monocytogenes* can use environmental flavins to shuttle electrons to outlying acceptors.

Integrating all of our insights into the roles of the components of the EET apparatus, we propose a molecular model of electron travel from intracellular NADH to membrane-confined quinone, then to extracellular flavoprotein (and/or other shuttles), and ultimately to a kinetically favourable terminal electron acceptor (Fig. 3d). To determine whether EET established a bona fide growth-supporting activity, we next screened a library of common microbial growth substrates and found that the inclusion of ferric iron or an electrode was required for anaerobic growth on the sugar alcohols xylitol and D-arabitol (Fig. 4a, Extended Data Fig. 7). Genes for aerobic respiration but not EET were essential for aerobic growth on xylitol, whereas this pattern was reversed under anaerobic conditions—that is, EET genes were essential and aerobic respiration genes dispensable (Fig. 4a, Extended Data Fig. 7). These data demonstrate that the distinct electron transport chains that segregate aerobic respiration and EET promote aerobic and anaerobic growth, respectively.

We next asked whether EET has a role in host colonization. Consistent with EET being dispensable for aerobic growth,

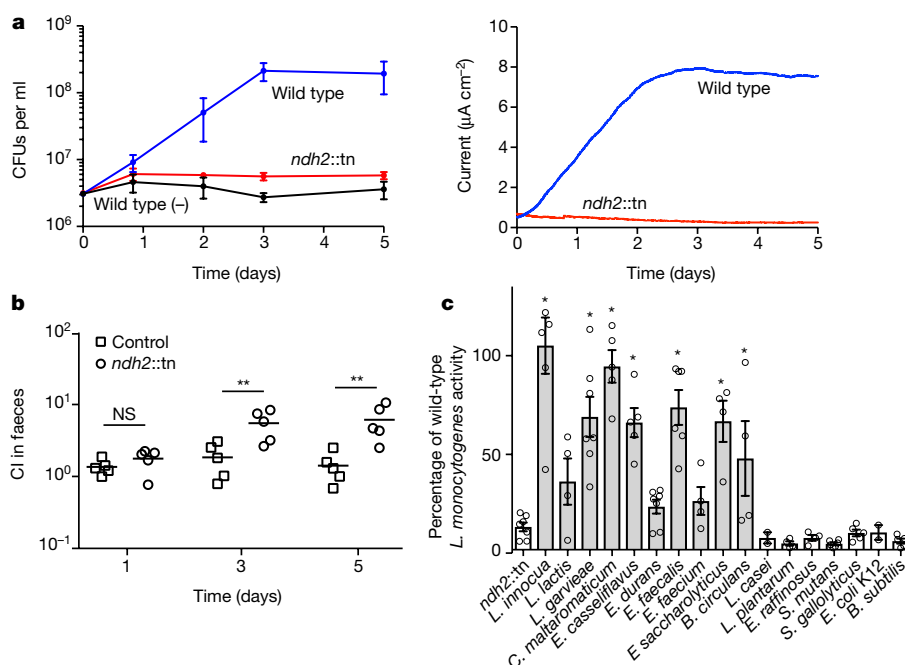


Fig. 4 | EET supports anaerobic growth, confers a competitive advantage in the intestinal lumen, and is active in multiple Firmicutes.

a, *L. monocytogenes* CFUs (left) and electric current (right) from chronoamperometry experiments conducted with xylitol growth medium. (–), control condition without an electrode. Results from three independent experiments are expressed as mean \pm s.e.m. **b**, Mice ($n = 5$) were fed bread inoculated with a 1:1 mixture of Δhly and $\Delta hly\ ndh2::tn$ *L. monocytogenes* strains. The competitive index (CI) at three time points after infection is indicated. Median values and statistically significant differences compared to a control that competed two Δhly strains are indicated; $**P = 0.01$, unpaired two-sided *t*-test. Results are representative of three independent experiments. **c**, Iron reductase activity in a panel of Firmicutes species, expressed as a percentage of wild-type *L. monocytogenes* activity. Results from at least three independent experiments ($n = 7$ for *ndh2::tn*, *Lactococcus garvieae* and *Enterococcus*

durans; $n = 6$ for *Listeria innocua*, *E. faecalis* and *Streptococcus mutans*; $n = 5$ for *Carnobacterium maltaromaticum*, *Enterococcus casseliflavus*, *Streptococcus gallolyticus* and *Bacillus subtilis*; $n = 4$ for *Lactococcus lactis*, *Enterococcus faecium*, *Enterococcus saccharolyticus*, *Bacillus circulans*, *Lactobacillus plantarum* and *Enterococcus raffinosus*; $n = 3$ for *Lactobacillus casei* and *E. coli* K12) are expressed as mean \pm s.e.m. Strains that statistically differ from *ndh2::tn* are indicated; $*P < 0.05$, ANOVA with Dunnett's post-test. Some members of Lactobacillales lack the ability to synthesize DHNA, the precursor for demethylmenaquinone biosynthesis, and require an exogenous source for quinone-dependent processes³⁶. Organisms with genes in the EET locus and *menC* (which catalyses an essential step in DHNA biosynthesis) are coloured grey. *L. casei*, *L. plantarum* and *E. raffinosus* contain genes for EET, but not *menC*. The remaining species lack genes in the EET locus.

EET-deficient mutants resembled wild-type *L. monocytogenes* in an intracellular macrophage growth assay and an intravenous infection model (Extended Data Fig. 8). Because anaerobic growth mechanisms are important for microbial proliferation within the intestinal lumen^{23,24}, we proposed that the food-borne pathogen might use EET in this context. Consistent with the hypothesis, the faecal burden of the *ndh2::tn* strain was decreased approximately sixfold in a streptomycin-pretreated model of *L. monocytogenes* intestinal colonization (Fig. 4b). These results thus suggest a role for EET within the dysbiotic gut and raise the possibility that EET constitutes a generally important metabolic activity within the mammalian gastrointestinal tract.

We next turned to the phylogenetic distribution of the genes responsible for EET. BLAST searches revealed that homologues of these genes are widespread in hundreds of species that span the Firmicutes phylum (Extended Data Fig. 9a, Supplementary Table 3). Many of these genes are likely to encode functional EET systems, as the identified locus is typically conserved, though noteworthy distinctions are evident in some genomes (Extended Data Fig. 9b). Microorganisms that possess a locus with EET genes adopt a wide range of different lifestyles, including within thermophilic (*Caldanaerobius* spp., *Thermoanaerobacter* spp. and so on) and halophilic (*Halolactibacillus* spp., *Halothermothrix* spp. and so on) habitats. Orthologues of the identified genes in the EET locus are found in a number of human pathogens (*Clostridium perfringens*, *Enterococcus faecalis*, *Streptococcus dysgalactiae* and so on), members of the human microbiota (*Clostridium* spp., *Enterococcus* spp., *Streptococcus* spp. and so on) and lactic acid bacteria that have

commercial applications in food fermentation or probiotics (*Lactococcus* spp., *Lactobacillus* spp., *Oenococcus* spp., *Tetragenococcus* spp. and so on) (Supplementary Table 3). The functionality of identified loci could explain previous reports of EET-like activity in a number of species^{25–35} and assays of ferric iron reductase activity of a panel of Firmicutes provided additional evidence that the presence of necessary genetic components correlates with EET activity (Fig. 4c).

In conclusion, our study reveals a novel electron transport chain that supports growth on extracellular electron acceptors. This mechanism lacks an elaborate multi-haem apparatus and, partly by taking advantage of the single-membrane architecture of the Gram-positive cell, is characterized by considerably fewer electron transfer steps than comparable systems in mineral-respiring Gram-negative bacteria¹. The genes identified in the EET locus are present in a wide-ranging group of microorganisms that occupy a diverse array of ecological niches. Defying conventional views of EET, this distinctive system is abundant in bacteria that prioritize fermentative metabolic strategies and reside in nutrient-rich environments, including the lactic acid bacteria. Within this context, environmental flavins may represent a feature of the ecological landscape that can be exploited to promote EET activity. These observations suggest that, rather than being a specialized process confined to mineral-respiring bacteria, the use of extracellular electron acceptors represents a fundamental facet of microbial metabolism that is relevant across diverse environments. In addition to obvious bioenergetic applications, the characterization of a flavin-based EET mechanism thus establishes further avenues for the study of electrochemical activities throughout the microbial world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0498-z>.

Received: 2 January 2018; Accepted: 3 August 2018;

Published online 12 September 2018.

- Shi, L. et al. Extracellular electron transfer mechanisms between microorganisms and minerals. *Nat. Rev. Microbiol.* **14**, 651–662 (2016).
- Myers, C. R. & Nealson, K. H. Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor. *Science* **240**, 1319–1321 (1988).
- Lovley, D. R. & Phillips, E. J. Novel mode of microbial energy metabolism: organic carbon oxidation coupled to dissimilatory reduction of iron or manganese. *Appl. Environ. Microbiol.* **54**, 1472–1480 (1988).
- Carlson, H. K. et al. Surface multiheme c-type cytochromes from *Thermincola potens* and implications for respiratory metal reduction by Gram-positive bacteria. *Proc. Natl Acad. Sci. USA* **109**, 1702–1707 (2012).
- Freitag, N. E., Port, G. C. & Miner, M. D. *Listeria monocytogenes* – from saprophyte to intracellular pathogen. *Nat. Rev. Microbiol.* **7**, 623–628 (2009).
- Deneer, H. G. & Boychuk, I. Reduction of ferric iron by *Listeria monocytogenes* and other species of *Listeria*. *Can. J. Microbiol.* **39**, 480–485 (1993).
- Kim, B. H., Kim, H. J., Hyun, M. S. & Park, D. H. Direct electrode reaction of Fe(III)-reducing bacterium, *Shewanella putrefaciens*. *J. Microbiol. Biotechnol.* **9**, 127–131 (1999).
- Marsili, E., Rollefson, J. B., Baron, D. B., Hozalski, R. M. & Bond, D. R. Microbial biofilm voltammetry: direct electrochemical characterization of catalytic electrode-attached biofilms. *Appl. Environ. Microbiol.* **74**, 7329–7337 (2008).
- Xu, S. J. Y. & El-Naggar, M. Y. Disentangling the roles of free and cytochrome-bound flavins in extracellular electron transport from *Shewanella oneidensis* MR-1. *Electrochim. Acta* **198**, 49–55 (2016).
- Karpowich, N. K., Song, J. M., Cocco, N. & Wang, D. N. ATP binding drives substrate capture in an ECF transporter by a release-and-catch mechanism. *Nat. Struct. Mol. Biol.* **22**, 565–571 (2015).
- Kerscher, S., Dröse, S., Zickermann, V. & Brandt, U. The three families of respiratory NADH dehydrogenases. *Results Probl. Cell Differ.* **45**, 185–222 (2008).
- Uden, G. & Bongaerts, J. Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim. Biophys. Acta* **1320**, 217–234 (1997).
- Bertsova, Y. V. et al. Alternative pyrimidine biosynthesis protein ApbE is a flavin transferase catalyzing covalent attachment of FMN to a threonine residue in bacterial flavoproteins. *J. Biol. Chem.* **288**, 14276–14286 (2013).
- Deka, R. K., Brautigam, C. A., Liu, W. Z., Tomchick, D. R. & Norgard, M. V. Evidence for posttranslational protein flavinylation in the syphilis spirochete *Treponema pallidum*: structural and biochemical insights from the catalytic core of a periplasmic flavin-trafficking protein. *MBio* **6**, e00519-15 (2015).
- Zückert, W. R. Secretion of bacterial lipoproteins: through the cytoplasmic membrane, the periplasm and beyond. *Biochim. Biophys. Acta* **1843**, 1509–1516 (2014).
- Glasser, N. R., Saunders, S. H. & Newman, D. K. The colorful world of extracellular electron shuttles. *Annu. Rev. Microbiol.* **71**, 731–751 (2017).
- Brutinel, E. D. & Gralnick, J. A. Shuttling happens: soluble flavin mediators of extracellular electron transfer in *Shewanella*. *Appl. Microbiol. Biotechnol.* **93**, 41–48 (2012).
- Marsili, E. et al. *Shewanella* secretes flavins that mediate extracellular electron transfer. *Proc. Natl Acad. Sci. USA* **105**, 3968–3973 (2008).
- von Canstein, H., Ogawa, J., Shimizu, S. & Lloyd, J. R. Secretion of flavins by *Shewanella* species and their role in extracellular electron transfer. *Appl. Environ. Microbiol.* **74**, 615–623 (2008).
- Kotloski, N. J. & Gralnick, J. A. Flavin electron shuttles dominate extracellular electron transfer by *Shewanella oneidensis*. *MBio* **4**, e00553-12 (2013).
- Powers, H. J. Riboflavin (vitamin B-2) and health. *Am. J. Clin. Nutr.* **77**, 1352–1360 (2003).
- Hühner, J., Ingles-Prieto, Á., Neusüß, C., Lämmerhofer, M. & Janovjak, H. Quantification of riboflavin, flavin mononucleotide, and flavin adenine dinucleotide in mammalian model cells by CE with LED-induced fluorescence detection. *Electrophoresis* **36**, 518–525 (2015).
- Winter, S. E. et al. Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* **467**, 426–429 (2010).
- Winter, S. E. et al. Host-derived nitrate boosts growth of *E. coli* in the inflamed gut. *Science* **339**, 708–711 (2013).
- Slobodkin, A. I. et al. Dissimilatory reduction of Fe(III) by thermophilic bacteria and archaea in deep subsurface petroleum reservoirs of western Siberia. *Curr. Microbiol.* **39**, 99–102 (1999).
- Roh, Y. et al. Isolation and characterization of metal-reducing thermoanaerobacter strains from deep subsurface environments of the Piceance Basin, Colorado. *Appl. Environ. Microbiol.* **68**, 6013–6020 (2002).
- Ogg, C. D. & Patel, B. K. *Fervidicola ferrireducens* gen. nov., sp. nov., a thermophilic anaerobic bacterium from geothermal waters of the Great Artesian Basin, Australia. *Int. J. Syst. Evol. Microbiol.* **59**, 1100–1107 (2009).
- Ogg, C. D. & Patel, B. K. *Thermotalea metallivorans* gen. nov., sp. nov., a thermophilic, anaerobic bacterium from the Great Artesian Basin of Australia aquifer. *Int. J. Syst. Evol. Microbiol.* **59**, 964–971 (2009).
- Ogg, C. D., Greene, A. C. & Patel, B. K. *Thermovenabulum gondwanense* sp. nov., a thermophilic anaerobic Fe(III)-reducing bacterium isolated from microbial mats thriving in a Great Artesian Basin bore runoff channel. *Int. J. Syst. Evol. Microbiol.* **60**, 1079–1084 (2010).
- Ogg, C. D. & Patel, B. K. *Fervidicella metallireducens* gen. nov., sp. nov., a thermophilic, anaerobic bacterium from geothermal waters. *Int. J. Syst. Evol. Microbiol.* **60**, 1394–1400 (2010).
- Masuda, M., Freguia, S., Wang, Y. F., Tsujimura, S. & Kano, K. Flavins contained in yeast extract are exploited for anodic electron transfer by *Lactococcus lactis*. *Bioelectrochemistry* **78**, 173–175 (2010).
- Zhang, E., Cai, Y., Luo, Y. & Piao, Z. Riboflavin-shuttled extracellular electron transfer from *Enterococcus faecalis* to electrodes in microbial fuel cells. *Can. J. Microbiol.* **60**, 753–759 (2014).
- Dong, Y. et al. *Orenia metallireducens* sp. nov. strain Z6, a novel metal-reducing member of the phylum Firmicutes from the deep subsurface. *Appl. Environ. Microbiol.* **82**, 6440–6453 (2016).
- Keogh, D. et al. Extracellular electron transfer powers *Enterococcus faecalis* biofilm metabolism. *MBio* **9**, e00626-17 (2018).
- Pankratova, G., Leech, D., Gorton, L. & Hederstedt, L. Extracellular electron transfer by the Gram-positive bacterium *Enterococcus faecalis*. *Biochemistry* **57**, 4597–4603 (2018).
- Pedersen, M. B., Gaudu, P., Lechardeur, D., Petit, M. A. & Gruss, A. Aerobic respiration metabolism in lactic acid bacteria and uses in biotechnology. *Annu. Rev. Food Sci. Technol.* **3**, 37–58 (2012).

Acknowledgements We thank G. Chen, J.-D. Sauer, E. Stevens, M. Marco and N. Freitag for providing bacterial strains; H. Carlson, A. Williamson and J. Coates for helpful feedback; and N. Garelis for experimental assistance. Research reported in this publication was supported by funding from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (F32AI136389 to S.H.L., 1P01 AI063302 to D.A.P., and 1R01 AI27655 to D.A.P.), the Office of Naval Research (N0001417WX01603 to C.M.A.-F.), and the China Scholarship Council (no. 2016060900098 to L.S.). A mass spectrometer used in this study was purchased with NIH support (grant 1S10OD020062-01). Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Reviewer information *Nature* thanks N. Freitag, J. Gralnick, K. Nealson and G. Reguera for their contribution to the peer review of this work.

Author contributions S.H.L., A.T.I., C.M.A.-F. and D.A.P. designed the study. S.H.L., L.S. and J.A.C. performed electrochemical experiments. S.H.L. and A.T.I. performed mass spectrometric experiments. S.H.L., A.L. and R.R.-L. performed microbiological and biochemical experiments. S.H.L. and D.A.P. wrote the manuscript.

Competing interests D.A.P. has a consulting relationship with and a financial interest in Aduro Biotech; both he and the company stand to benefit from the commercialization of this research.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0498-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0498-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.A.P. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

***L. monocytogenes* strains and growth conditions.** All *L. monocytogenes* strains used in this study were derived from wild-type 10403S (Supplementary Table 4). Transduction methods were used to introduce transposons into distinct genetic backgrounds, as previously described^{37,38}. *L. monocytogenes* cells were grown at 37 °C and spectrophotometrically measured by optical density at a wavelength of 600 nm (OD₆₀₀). Anaerobic conditions were achieved with the BD GasPak EZ pouch system or an anaerobic chamber (Coy Laboratory Products) with an environment of 2% H₂ balanced in N₂.

Filter-sterilized brain–heart infusion medium (Difco) or variants of chemically defined *Listeria* synthetic medium (LSM)³⁹ were used in all studies. Aerobic respiration medium replaced the glucose in LSM with 50 mM glycerol. The requirement of an electron acceptor to support *L. monocytogenes* growth on xylitol was identified by comparing aerobic versus anaerobic (absent an alternative electron acceptor) growth on carbon sources, using PM1 and PM2A plates of the Phenotype MicroArray (Biolog). ‘Xylitol medium’ replaced the glucose in LSM with 50 mM xylitol.

Gene name assignment. The identified EET locus is widely conserved in *L. monocytogenes* isolates and encompasses the genes *lmrg_02179–lmrg_02186* in *L. monocytogenes* 10403S (which correspond to *lmo2634–lmo2641* in *L. monocytogenes* EGD-e). Identified genes in the EET locus were assigned *dmk* or *fnn* prefixes based on putative roles in demethylmenaquinone biosynthesis or PplA FMNylation, respectively. The *eet* prefix was assigned to the remaining genes, which at present lack high-confidence functional assignments. The only previously named gene, *pplA*, was so-called on the basis of the role of its cleaved signal peptide as a signalling pheromone⁴⁰ (a function that seems to be unrelated to the mature protein).

Bioelectrochemical characterization and measurements. Chronoamperometry and cyclic voltammetry were carried out using a Bio-Logic Science Instruments potentiostat model VSP-300. All measurements were performed using double chamber electrochemical cells (Extended Data Fig. 1a) and consisted of an Ag/AgCl reference electrode (CH Instruments), a Pt wire counter electrode (Alfa Aesar), and a 6.35-mm-thick graphite felt working electrode with a 16-mm radius (Alfa Aesar).

Electrochemical cells were prepared with 120 ml of modified LSM (containing 0.8 μM FMN as the sole flavin) and an open circuit potential was performed in the absence of bacteria. Once the current stabilized, the electrochemical cell was inoculated to a final OD₆₀₀ of ~0.1. The medium in the electrochemical chamber was mixed with a magnetic stir bar for the course of the experiment. For current acquisition, the applied potential was set at +0.4 V versus Ag/AgCl. To maintain anaerobic conditions, electrochemical cells were continuously purged with N₂ gas. Cyclic voltammetry measurements in the potential region of –0.8 to +0.4 V versus Ag/AgCl and a scan rate of 10 mV s^{–1} were conducted immediately before inoculation and 3 h later. Electric currents are reported as a function of the geometric surface area of the electrode. To test the effect of flavins on electrochemical activity, FMN was injected into the *L. monocytogenes*-inoculated electrochemical chamber to a final concentration of 1 μM.

For *S. oneidensis* experiments, the glucose in LSM was replaced with sodium lactate and *S. oneidensis* was inoculated to an OD₆₀₀ of 0.1. Growth-supporting *L. monocytogenes* experiments on xylitol medium were conducted in a similar fashion, but the electrochemical cell was inoculated to an OD₆₀₀ of ~0.002 and the medium from the electrochemical chamber was sampled at regular intervals for the enumeration of CFUs.

Screen of mutants with diminished ferric iron reductase activity. A previously described method was adapted to screen for *L. monocytogenes* mutants with diminished ferric iron reductase activity⁶. Approximately 250 CFUs per plate of a pooled *himar1* transposon library, generated as previously described³⁸, were grown on brain–heart infusion agar supplemented with 0.1 mg/ml ferric ammonium citrate. After 24 h at 37 °C, plates were removed from the incubator and a 10-ml overlay (0.8% agarose and 2 mM ferrozine) was applied. Colorimetric change resulting from ferrozine binding to Fe²⁺ was visually tracked for ~10 min. Colonies with diminished colorimetric change were selected and the location of the transposon insertion identified by Sanger sequencing, as previously described⁴¹.

Ferrozine assay of ferric iron reductase activity. *L. monocytogenes* cells grown to mid-log phase were washed twice, normalized to an OD₆₀₀ of 0.5, and resuspended in fresh medium supplemented with 4 mM ferrozine. Experiments were initiated by adding 100 μl of cells to an equivalent volume of 50 mM ferric ammonium citrate or ferric (hydr)oxide and were conducted in triplicate at 37 °C in 96-well format using a plate reader. OD₅₆₂ measurements were made every 30 s for up to an hour. Maximal rates (typically over 2 min) calculated from a Fe²⁺ standard curve are reported. Assays were generally performed in LSM, with glucose serving as the electron donor. However, because some of the respiratory mutants grew poorly in these conditions, these strains were assayed in brain–heart infusion medium (with glucose remaining as the electron donor). For FAD complementation studies, before washing steps, strains grown to mid-log were split and—after adding

0.5 mM FAD to one aliquot—incubated for 1 h at 37 °C. To test the effect of flavins, riboflavin, FMN or FAD was titrated into cells resuspended in a LSM base that lacked flavins.

To prepare other species (detailed in Supplementary Table 4) for the ferric iron reductase assay, cells were grown anaerobically in brain–heart infusion medium for 36 h. Sub-cultures in brain–heart infusion medium supplemented with 25 mM ferric ammonium citrate were then grown to mid-log phase. Cells were washed twice, resuspended in fresh brain–heart infusion medium and cell densities were normalized to wild-type *L. monocytogenes*. Next, ferrozine was added to a final concentration of 2 mM and 100 μl of cells were dispensed in a 96-well plate. The experiment was initiated by adding 100 μl of brain–heart infusion medium supplemented with 10 mM ferric ammonium citrate and OD₅₆₂ measurements were made as described for the *L. monocytogenes* ferric iron reductase assay.

***L. monocytogenes* growth on xylitol and ferric iron.** To test electron acceptor usage capabilities, xylitol medium was inoculated with *L. monocytogenes* and incubated at 25 °C in an anaerobic chamber. Conditions testing putative electron acceptors contained 50 mM ferric ammonium citrate or ferric (hydr)oxide, prepared as previously described⁴². For the ferric ammonium citrate experiments, 50 mM sodium citrate was included in the control condition that lacked ferric ammonium citrate and CFUs were enumerated following overnight incubation in a 96-well plate (Greiner Bio-One). Ferric (hydr)oxide experiments were conducted in a 6-well plate (Costar) and CFUs were enumerated 6 days after inoculation.

NAD⁺ and NADH measurements. *L. monocytogenes* cells grown overnight in LSM were washed and resuspended in 500 μl of medium. Cells were then split and 50 mM ferric ammonium citrate was added to one aliquot. To test aerobic conditions, 14-ml tubes were placed in a shaking (200 r.p.m.) incubator. To achieve microaerophilic conditions, the headspace in the tube was purged with argon gas and the tightly capped tube was placed in a stationary incubator. After 1.5 h at 37 °C, bacteria were collected by centrifugation, resuspended in PBS and lysed by vortexing with 0.1-mm-diameter zirconia–silica beads. NAD⁺ and NADH measurements were performed using the NAD/NADH-Glo Assay (Promega).

Assay of FmnB FMN transferase activity. Constructs of *fmnB* and *pplA* that truncated the signal peptide were subcloned into the pMCSG58 vector. Protein overexpression and purification followed previously described protocols⁴³. Purified PplA and FmnB were incubated overnight at a 10:1 molar ratio in assay buffer (0.5 M NaCl and 10 mM Tris, pH 8.3) with putative flavin substrates. Because homologous FMN transferases require a magnesium cofactor¹³, the effect of the chelator EDTA on activity was tested. Samples were analysed by SDS–PAGE and protein bands with covalent flavin modifications were visualized by UV illumination.

To identify the basis of post-translational modifications, intact protein mass measurements of PplA were made using a Synapt G2-Si mass spectrometer that was equipped with an electrospray ionization source and a C₄ protein ionKey (inner diameter: 150 μm, length: 50 mm, particle size: 1.7 μm), and connected in-line with an Acquity M-class ultra-performance liquid chromatography system (UPLC; Waters). Acetonitrile, formic acid (Fisher Optima grade, 99.9%) and water purified to a resistivity of 18.2 MΩ-cm (at 25 °C) using a Milli-Q Gradient ultrapure water purification system (Millipore) were used to prepare mobile phase solvents. Solvent A was 99.9% water/0.1% formic acid and solvent B was 99.9% acetonitrile/0.1% formic acid (v/v). The elution program consisted of a linear gradient from 1% to 10% B (v/v) over 1 min, a linear gradient from 10% to 90% B over 4 min, isocratic flow at 90% B for 5 min, a linear gradient from 90% to 1% B over 2 min, and isocratic flow at 1% B for 18 min, at a flow rate of 2 μl/min. The ionKey column and the autosampler compartment were maintained at 40 °C and 6 °C, respectively. Mass spectra were acquired in the positive ion mode and continuum format, operating the time-of-flight analyser in resolution mode, with a scan time of 0.5 s, over the range *m/z* = 400 to 5,000. Mass spectral deconvolution was performed using ProMass software (version 2.5 SR-1, Novatia).

***L. monocytogenes* protein trypsinization.** One millilitre of *L. monocytogenes* cells grown in LSM to mid-log phase was washed, resuspended in 100 μl of 100 mM NH₄HCO₃ (pH 7.5), and incubated at 100 °C for 10 min. Cells were lysed by bead beating for 15 min at 4 °C. RapiGest SF (Waters) was added to lysed cells at a final concentration of 0.1% and the sample was incubated at 100 °C for 5 min. After adding 5 μl of 100 mM dithiothreitol, samples were incubated at 58 °C for 30 min. Next, 15 μl of 100 mM iodoacetamide was added and samples were incubated for an additional 30 min. Samples were then digested overnight with 10 μl Trypsin Gold (Promega). The following morning, 10 μl of 5% trifluoroacetic acid was added and samples were incubated at 37 °C for 90 min. Samples were centrifuged for 30 min to remove hydrolysed RapiGest, and supernatant was collected.

***L. monocytogenes* intracellular growth assays.** Bone-marrow-derived macrophages prepared from 6- to 8-week-old female mice were plated overnight on coverslips and infected with *L. monocytogenes* strains at a multiplicity of infection of 0.1. Macrophage monolayers were washed with PBS and fresh medium was added thirty minutes after infection. At 1 h post-infection, 50 μg/ml gentamicin was added to kill extracellular bacteria. To enumerate *L. monocytogenes* CFUs,

macrophages were lysed by transferring coverslips to 10 ml of water, as previously described⁴⁴.

***L. monocytogenes* intravenous infections.** Eight-week-old female C57BL/6 mice (The Jackson Laboratory) were infected with 1×10^5 CFUs in 200 μ l of PBS by tail vein injection. Forty-eight hours post-infection, spleens and livers were collected, homogenized and plated for the enumeration of CFUs.

***L. monocytogenes* oral infections.** Previously described models of *L. monocytogenes* oral infection were adapted to address the role of EET in the intestinal lumen^{45,46}. Prior to infection, 5 mg/ml of streptomycin sulfate was added to the drinking water of 8-week-old female C57BL/6 mice (The Jackson Laboratory). After 24 h, mice were transferred to fresh cages and chow was removed to initiate an overnight fast. Forty-eight hours after streptomycin addition to the water, mice were isolated, fed a small piece of bread with 3 μ l of butter and an inoculum with 10^8 CFUs of *L. monocytogenes*, and returned to cages containing standard drinking water and chow. To confine *L. monocytogenes* to the intestinal lumen, a Δ hly parental strain (which has greatly reduced intracellular growth and spread) was used in these experiments. Inoculums were prepared with a 1:1 ratio of Δ hly and an erythromycin-resistant Δ hly strain (Δ hly *erm*^R, derived as previously described⁴⁷) or Δ hly and Δ hly *ndh2::tn*. Following infection, stools were collected, homogenized and dilutions were plated. Because total parental strain CFUs did not statistically differ between conditions, results are simply reported as a competitive index (that is, the ratio of streptomycin to erythromycin-resistant CFUs). These studies were carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All protocols were reviewed and approved by the Animal Care and Use Committee at the University of California, Berkeley (AUP-2016-05-8811).

Identification of protein substrates of FmnB. Wild-type and *fmnB::tn* strains grown in LSM were prepared for proteomic analysis as described in '*L. monocytogenes* protein trypsinization'. Peptides with >50% FMNylated peptide relative ion abundance in the wild-type sample and <5% in the *fmnB::tn* sample were identified using Progenesis QI for Proteomics software (version 4.0, Waters) and validated by manual inspection of the data. To address the FMNylation status of PplA, *ribU::tn* and *fmnA::tn* mutants were prepared in the same manner.

Trypsin-shaving analysis of surface-associated proteins. Trypsin-shaving experiments were adapted from a previously described method⁴. Cells grown in brain-heart infusion medium were washed twice and resuspended in a shaving buffer (1 M sucrose + 1 mM HEPES, pH 7). Lysozyme from chicken egg white (Sigma) was added to a concentration of 0.1 mg/ml. Cells were incubated at 37 °C for 60 min and released surface-associated components were separated from the cell by centrifugation. The supernatant (surface-associated protein fraction) was dialysed overnight in digestion buffer (100 mM NH₄HCO₃, pH 7.5) and the pellet (total protein fraction) was resuspended in digestion buffer. Samples were prepared for proteomic experiments as described in '*L. monocytogenes* protein trypsinization'. A label-free relative quantification approach^{48,49} implemented in Progenesis QI for Proteomics software (version 4.0, Waters) identified proteins that were disproportionately abundant in the surface-associated fraction.

Liquid chromatography-mass spectrometry analysis of trypsin-digested proteins. Samples of trypsin-digested proteins were analysed in triplicate using the Acquity M-class UPLC and Synapt G2-Si mass spectrometer, as follows. The mass spectrometer was equipped with a nano-electrospray ionization source that was connected in-line with the UPLC. The UPLC was equipped with trapping (Symmetry C18, inner diameter: 180 μ m, length: 20 mm, particle size: 5 μ m) and analytical (HSS T3, inner diameter: 75 μ m, length: 250 mm, particle size: 1.8 μ m, Waters) columns. Solvent A was 99.9% water/0.1% formic acid and solvent B was 99.9% acetonitrile/0.1% formic acid (v/v). The elution program consisted of a linear gradient from 1% to 10% B (v/v) over 2 min, a linear gradient from 10% to 35% B over 90 min, a linear gradient from 35% to 90% B over 1 min, isocratic flow at 90% B for 6 min, a linear gradient from 90% to 1% B over 1 min, and isocratic flow at 1% B for 20 min, at a flow rate of 300 nl/min. The column and autosampler compartments were maintained at 35 °C and 6 °C, respectively. Ion mobility-enabled HD-MS^E data^{50,51} were acquired in the positive ion mode and continuum format, operating the time-of-flight analyser in resolution mode, with a scan time of 0.5 s, over the range *m/z* = 50 to 2,000. An optimized wave velocity of 850 m/s was used for the travelling wave ion mobility cell. Collision-induced dissociation was performed in the ion transfer cell with a collision energy ramp from 30 to 78 V.

Data acquisition was controlled using MassLynx software (version 4.1), and tryptic peptides were identified using Progenesis QI for Proteomics software (version 4.0, Waters).

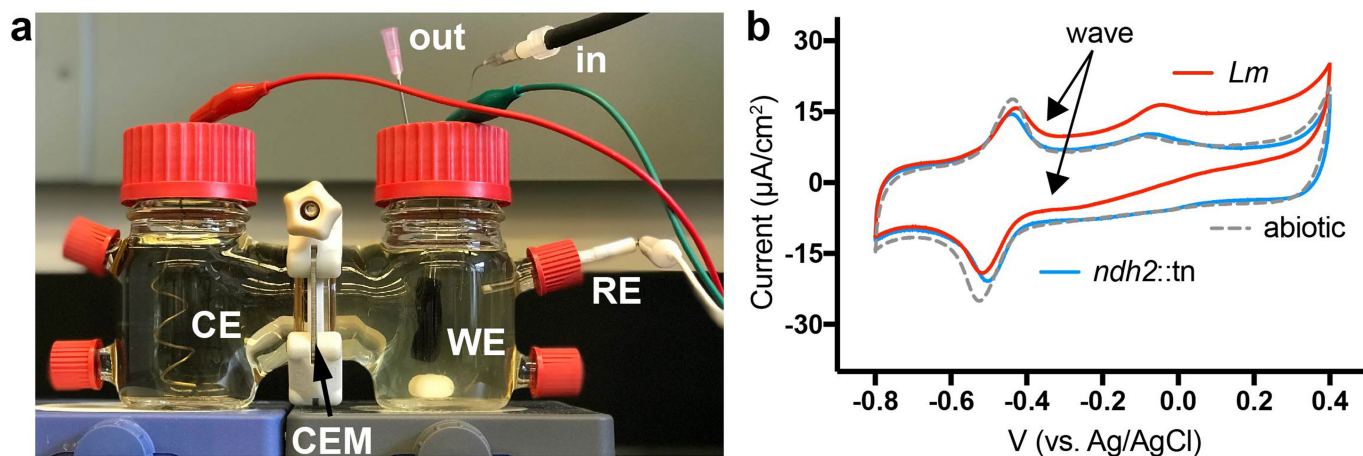
Bioinformatics analysis of identified genes in the EET locus. Ndh2 homologues were identified by searching the sequence of the unique C-terminal domain of Ndh2 on the PSI-BLAST server⁵². To perform a phylogenetic analysis, representative homologues were selected and aligned by ClustalW⁵³. The maximum likelihood method was used to infer the evolutionary history of identified sequences in Mega 7.0.26 and confidence limits of branch points were estimated by 1,000 bootstrap replications^{54,55}. The information about EET genetic loci summarized in Supplementary Table 3 was acquired by analysing genomic context of identified genes in the PATRIC 3.5.1 database (<https://www.patricbrc.org>).

Statistics and reproducibility. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment. Statistical analyses were performed in Prism 5 for Mac OS X (GraphPad Software) and Progenesis QI for Proteomics version 4.0.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

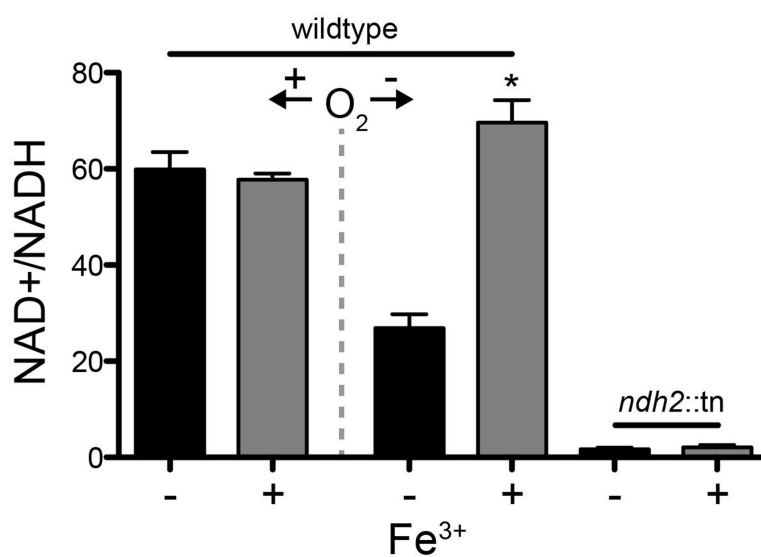
Data availability. The datasets generated during the current study are available from the corresponding author on reasonable request.

37. Hodgson, D. A. Generalized transduction of serotype 1/2 and serotype 4b strains of *Listeria monocytogenes*. *Mol. Microbiol.* **35**, 312–323 (2000).
38. Zemansky, J. et al. Development of a mariner-based transposon and identification of *Listeria monocytogenes* determinants, including the peptidyl-prolyl isomerase PrsA2, that contribute to its hemolytic phenotype. *J. Bacteriol.* **191**, 3950–3964 (2009).
39. Whiteley, A. T., Pollock, A. J. & Portnoy, D. A. The PAMP c-di-AMP is essential for *Listeria monocytogenes* growth in rich but not minimal media due to a toxic increase in (p)ppGpp. *Cell Host Microbe* **17**, 788–798 (2015).
40. Yayarath, B., Alonzo, F. III & Freitag, N. E. Identification of a peptide-pheromone that enhances *Listeria monocytogenes* escape from host cell vacuoles. *PLoS Pathog.* **11**, e1004707 (2015).
41. Burke, T. P. et al. *Listeria monocytogenes* is resistant to lysozyme through the regulation, not the acquisition, of cell wall-modifying enzymes. *J. Bacteriol.* **196**, 3756–3767 (2014).
42. Lovley, D. R. & Phillips, E. J. Organic matter mineralization with reduction of ferric iron in anaerobic sediments. *Appl. Environ. Microbiol.* **51**, 683–689 (1986).
43. Light, S. H., Cahoon, L. A., Halavaty, A. S., Freitag, N. E. & Anderson, W. F. Structure to function of an α -glucan metabolic pathway that promotes *Listeria monocytogenes* pathogenesis. *Nat. Microbiol.* **2**, 16202 (2016).
44. Portnoy, D. A., Jacks, P. S. & Hinrichs, D. J. Role of hemolysin for the intracellular growth of *Listeria monocytogenes*. *J. Exp. Med.* **167**, 1459–1471 (1988).
45. Bou Ghanem, E. N. et al. InlA promotes dissemination of *Listeria monocytogenes* to the mesenteric lymph nodes during food borne infection of mice. *PLoS Pathog.* **8**, e1003015 (2012).
46. Becattini, S. et al. Commensal microbes provide first line defense against *Listeria monocytogenes* infection. *J. Exp. Med.* **214**, 1973–1989 (2017).
47. Auerbuch, V., Lenz, L. L. & Portnoy, D. A. Development of a competitive index assay to evaluate the virulence of *Listeria monocytogenes* actA mutants during primary and secondary infection of mice. *Infect. Immun.* **69**, 5953–5957 (2001).
48. Neilson, K. A. et al. Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* **11**, 535–553 (2011).
49. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for label-free peptide quantification. *Mol. Cell. Proteomics* **12**, 549–556 (2013).
50. Plumb, R. S. et al. UPLC/MS^E: a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom.* **20**, 1989–1994 (2006).
51. Shliaha, P. V., Bond, N. J., Gatto, L. & Lilley, K. S. Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *J. Proteome Res.* **12**, 2323–2339 (2013).
52. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
53. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
54. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282 (1992).
55. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).



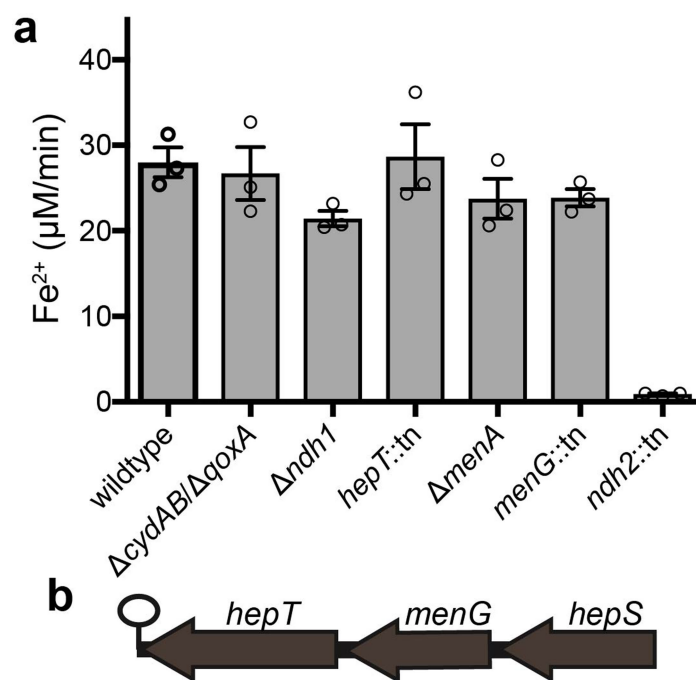
Extended Data Fig. 1 | Electrochemical analyses of *L. monocytogenes*.
a, The double chamber cell used for electrochemical experiments. CE, counter electrode; CEM, cation exchange membrane; RE, reference electrode; WE, working electrode. Inlets and outlets for N_2 gas are

labelled. **b**, Cyclic voltammograms of wild-type and *ndh2::tn* strains of *L. monocytogenes*. 'Abiotic' refers to an uninoculated control. Arrows highlight the initiation of the catalytic wave. Results are representative of three independent experiments.



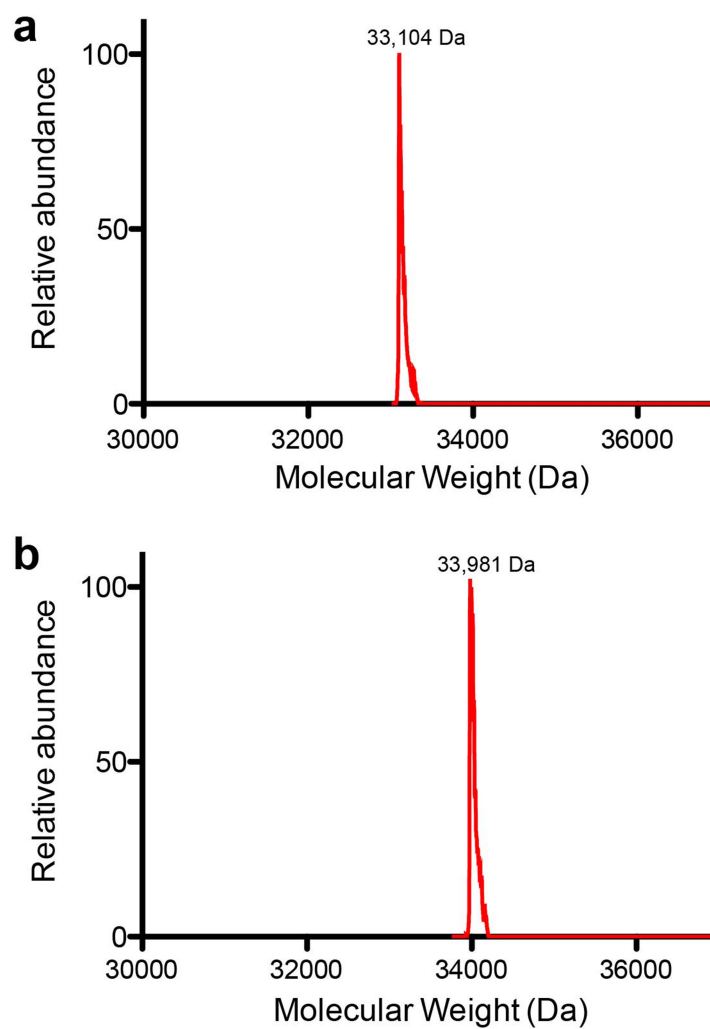
Extended Data Fig. 2 | EET activity maintains cellular redox homeostasis. Ratio of NAD⁺ to NADH in wild-type and *ndh2::tn* strains supplemented with ferric ammonium citrate under aerobic or microaerophilic conditions. Results from three independent experiments

are expressed as mean \pm s.e.m. A statistically significant difference between microaerophilic cells incubated with or without iron is indicated; * $P=0.0015$, unpaired two-sided t -test.



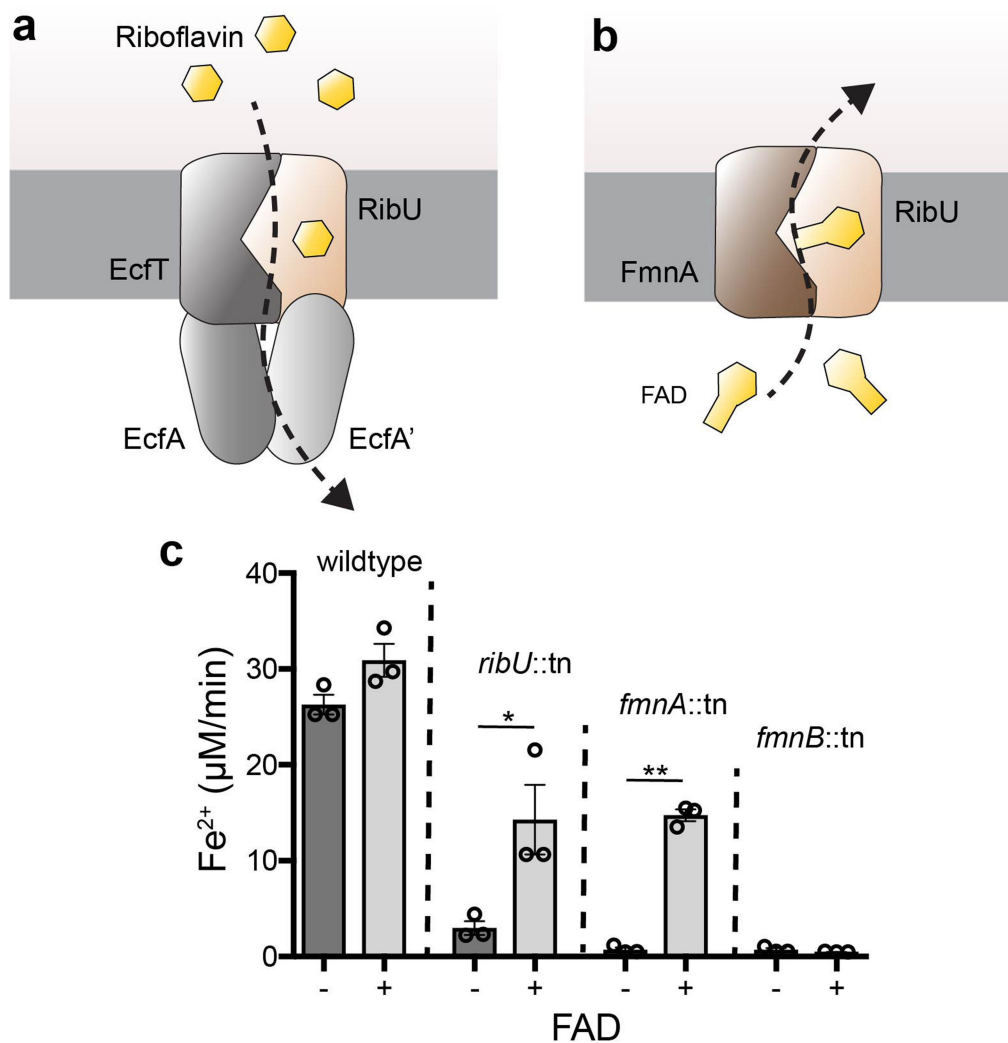
Extended Data Fig. 3 | Evidence that a distinct menaquinone derivative functions in aerobic respiration. a, Ferric iron reductase activity of mutants described in Fig. 2 demonstrates that genes essential for growth on aerobic respiration medium are dispensable for EET. Results from three independent experiments are expressed as mean

\pm s.e.m. **b,** The *L. monocytogenes* *hep* operon. Notably, *menG*—which encodes demethylmenaquinone transferase (the enzyme that converts demethylmenaquinone to menaquinone) (Fig. 2b)—neighbours the *hepT* and *hepS* genes, which function in quinone biosynthesis and are essential for aerobic respiration (Fig. 2c).



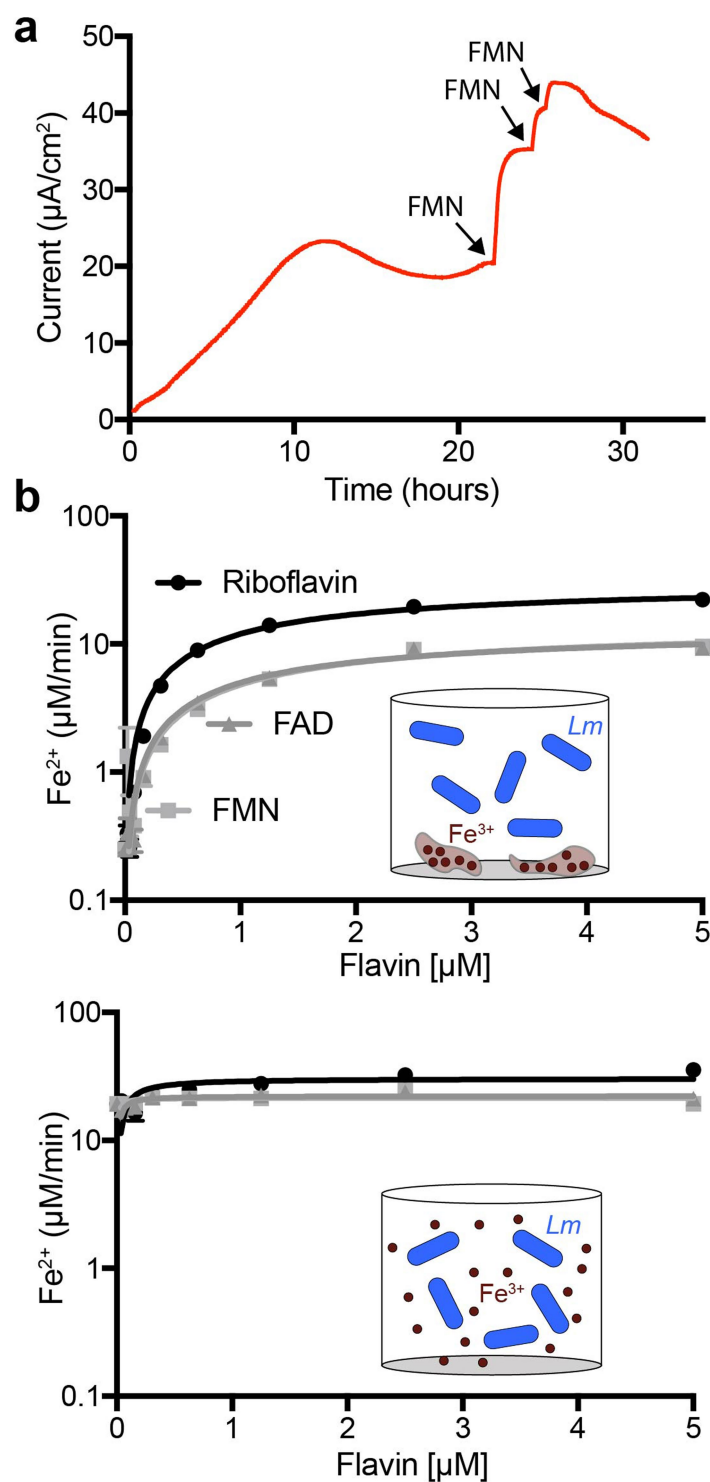
Extended Data Fig. 4 | Recombinant FmnB FMNylates PplA at two discrete sites. a, b, Deconvoluted mass spectra from a single experiment of recombinant PplA (a) and recombinant PplA incubated with

FAD + FmnB (b). The observed molecular weight change (877 Da) is consistent with two post-translational FMNylations (2×438.3 Da) on PplA.



Extended Data Fig. 5 | Proposed role of RibU and FmnA in FAD secretion. **a**, Simplified adaptation of a previously proposed model of *L. monocytogenes* riboflavin uptake through the RibU, EcfT, EcfA and EcfA' transporter¹⁰. According to this model, EcfT, EcfA and EcfA' couple ATP hydrolysis with conformational changes that result in substrate bound to RibU being released into the cytosol. **b**, On the basis of protein homology (FmnA shares 50% sequence identity with EcfT) and the expectation that extracellular FAD is required for FmnB to catalyse FMNylation of PplA, we propose that the FmnA interacts with RibU

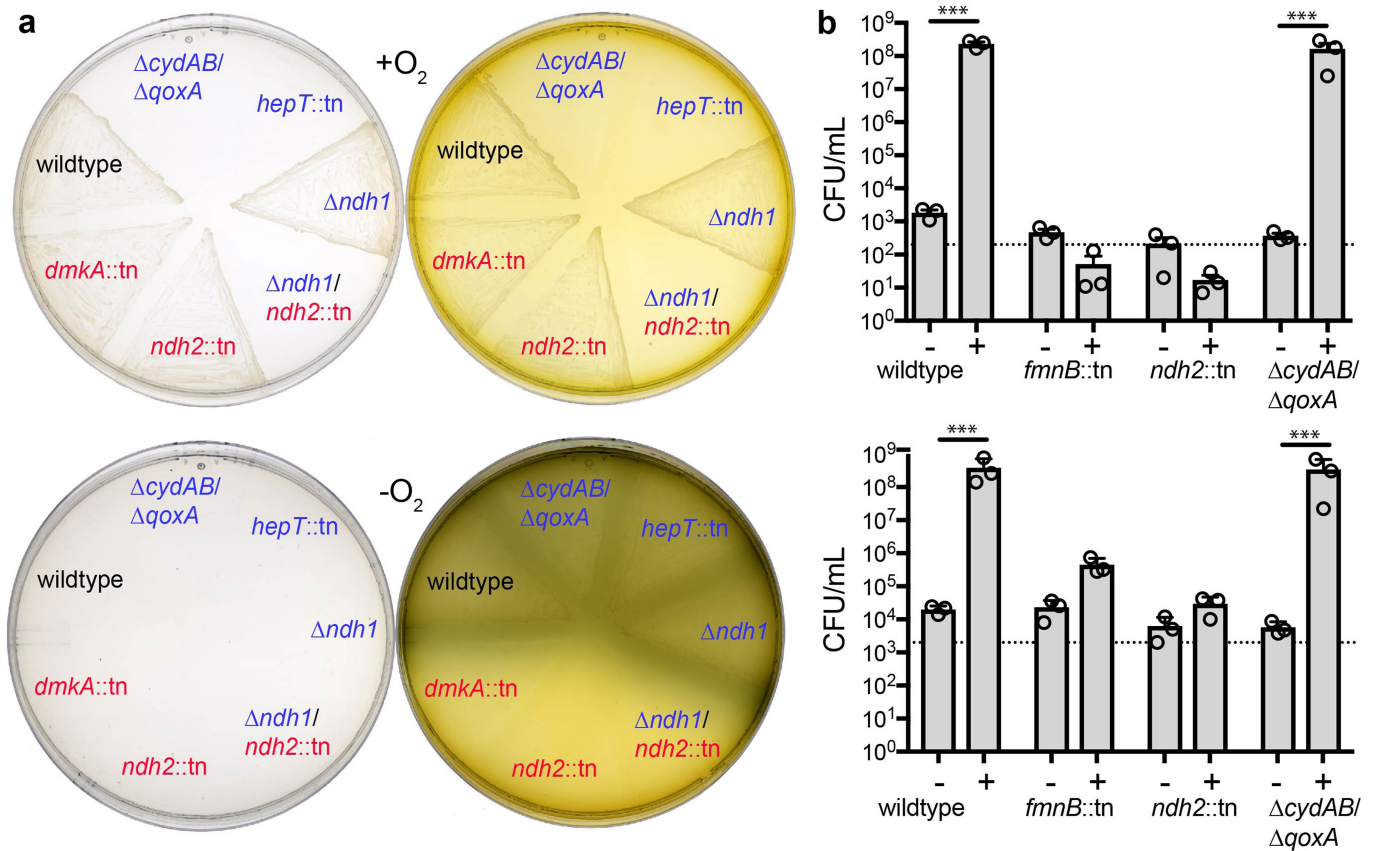
to promote FAD secretion. **c**, Ferric iron reductase activity of strains incubated with 0.5 mM FAD for 1 h. The ability of exogenous FAD to specifically rescue ferric iron reductase activity in the *fmnA::tn* and *ribU::tn* strains is consistent with FmnA and RibU functioning in FAD secretion. Results from three independent experiments are expressed as mean \pm s.e.m. Statistically significant differences between untreated and FAD-treated cells are indicated; * $P = 0.038$, ** $P < 0.0001$, unpaired two-sided *t*-test.



Extended Data Fig. 6 | Flavin shuttles promote EET activity.

a, Chronoamperometry results from *L. monocytogenes*-inoculated electrochemical reactors with $1\ \mu\text{M}$ FMN injections at the indicated time points. Results are representative of three independent experiments. **b**, The effect of flavins on *L. monocytogenes* (*Lm*) ferric iron reductase activity

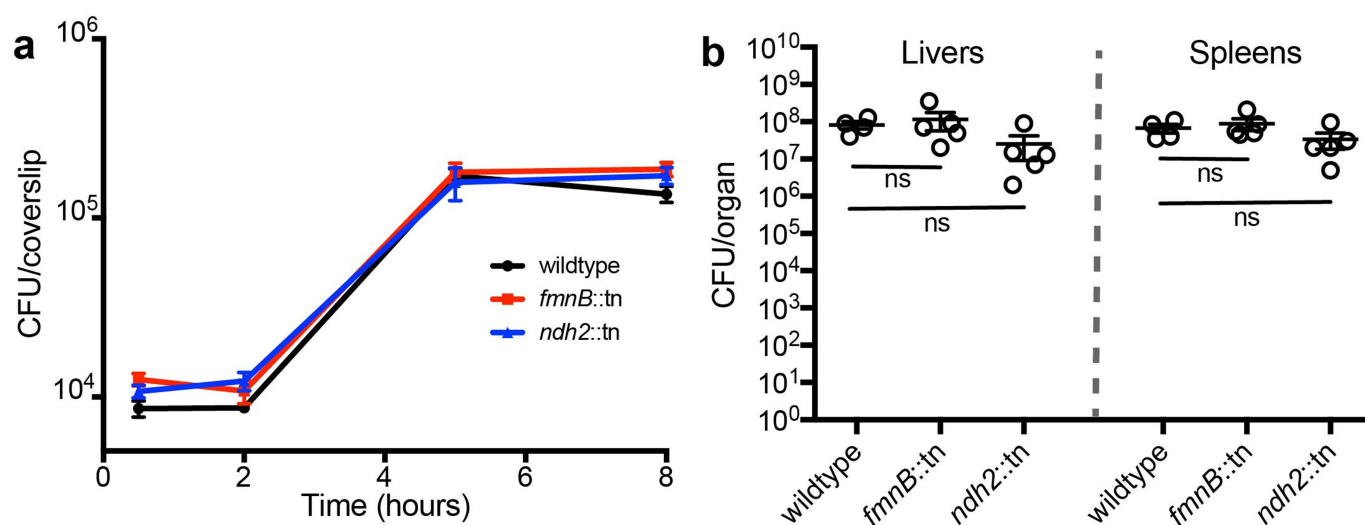
with insoluble ferric (hydr)oxide (top) and soluble ferric ammonium citrate (bottom). With insoluble substrate the local iron concentration for most cells is low, whereas with soluble substrate the concentration of iron in the direct vicinity of cells is high (insets). Results from three independent experiments are expressed as mean \pm s.e.m.



Extended Data Fig. 7 | EET supports anaerobic growth on ferric iron.

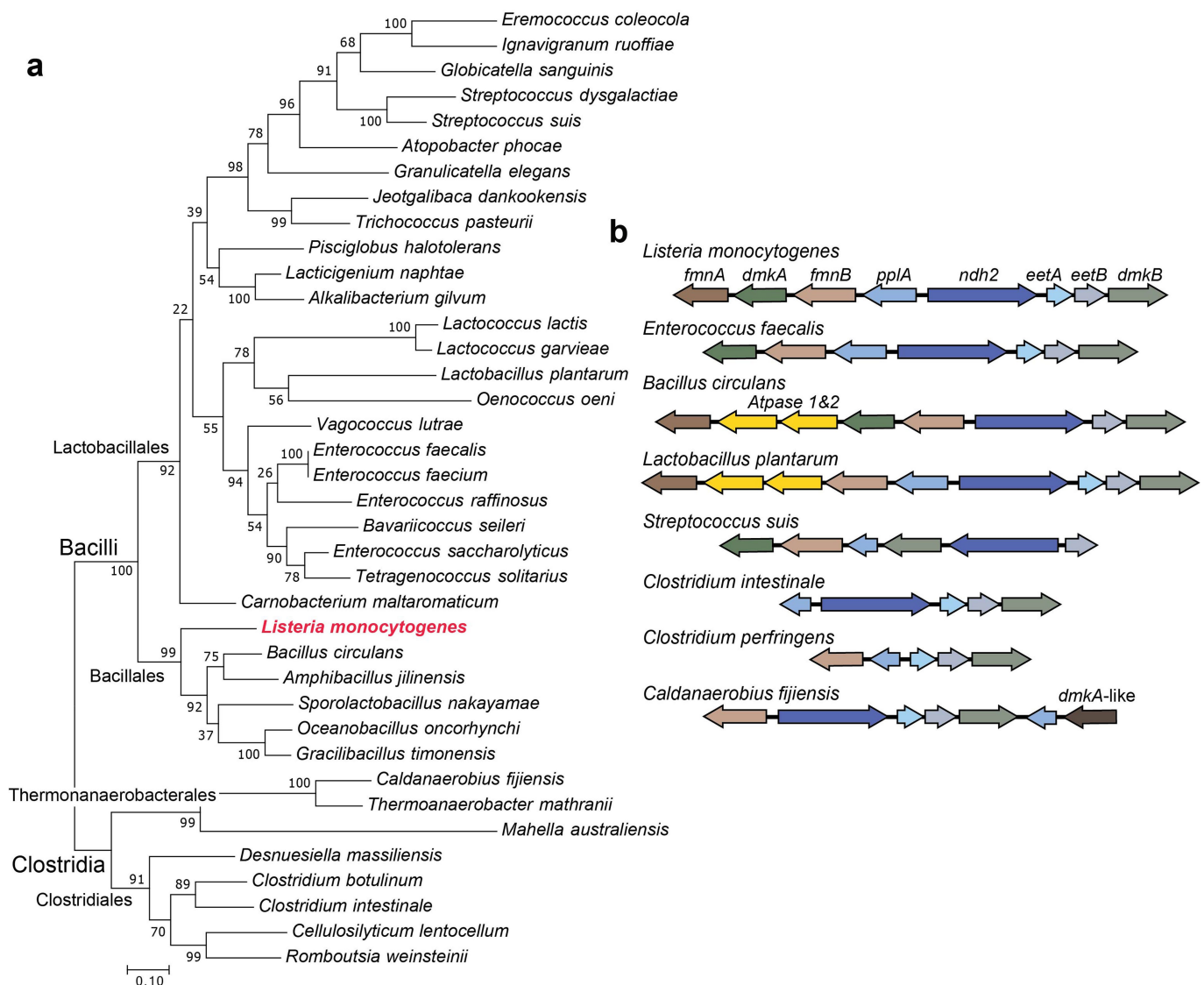
a, Growth following incubation of *L. monocytogenes* strains on xylitol medium without (left) or with (right) ferric iron under aerobic (top) or anaerobic (bottom) conditions. Results are representative of three independent experiments. Strain labels are coloured based on attributed deficiencies (Fig. 2d) in aerobic respiration (blue) or EET (red). Ndh1 and Ndh2 are probably functionally redundant under aerobic conditions, as a growth phenotype is only observed in the double mutant. Note the visual evidence of ferrous iron production in the agar adjoining anaerobically

growing cells. **b**, CFUs of *L. monocytogenes* strains anaerobically incubated in xylitol medium without (–) or with (+) ferric supplementation. Results for soluble ferric ammonium citrate (top) and insoluble ferric (hydr) oxide (bottom) are shown. Dashed lines denote the number of cells at the start of the experiment. Results from three independent experiments are expressed as mean \pm s.e.m. Statistically significant differences in the ferric iron-supplemented condition are noted; *** $P < 0.0001$, unpaired two-sided *t*-test.



Extended Data Fig. 8 | EET genes are dispensable for *L. monocytogenes* intracellular growth. a, Mouse bone-marrow-derived macrophages were infected with *L. monocytogenes*, and CFUs were enumerated at the indicated times. Results from three independent experiments are

expressed as mean \pm s.e.m. **b**, *L. monocytogenes* burdens in mouse organs ($n = 5$) 48 h after intravenous infection. Representative results from two independent experiments are expressed as median and s.e.



Extended Data Fig. 9 | Identified EET loci are widespread in the Firmicutes phylum. a, Phylogenetic tree constructed from select Ndh2 homologue sequences. A more comprehensive list of organisms that possess an EET locus is provided in Supplementary Table 3. Labels on the branches refer to the percentage of replicate trees that gave the depicted branch topology in a bootstrap test of 1,000 replicates. **b**, Distinct EET loci from select genomes are shown. Although the arrangement of genes varies, a locus with genes associated with EET is present in many genomes. Some loci contain ECF transporter ATPase subunits (homologous to

those depicted in Extended Data Fig. 5a) that probably function with RibU and FmnA subunits in flavin transport. The *dmkA*-like gene found in *Caldanaerobius fijiensis* (and other genomes) lacks homology to *dmkA*, but is annotated as catalysing the same reaction. The *pplA* variant in some genomes contains a single FMNylated domain (rather than two) and this property is indicated by a shorter arrow. A few bacteria (including *Lactococcus* spp.) lack a recognizable locus and distribute genes associated with EET throughout the genome.

Architecture of the TRPM2 channel and its activation mechanism by ADP-ribose and calcium

Yihe Huang¹, Paige A. Winkler¹, Weinan Sun^{2,3}, Wei Lü^{1*} & Juan Du^{1*}

Transient receptor potential melastatin 2 (TRPM2) is a calcium-permeable, non-selective cation channel that has an essential role in diverse physiological processes such as core body temperature regulation, immune response and apoptosis^{1–4}. TRPM2 is polymodal and can be activated by a wide range of stimuli^{1–7}, including temperature, oxidative stress and NAD⁺-related metabolites such as ADP-ribose (ADPR). Its activation results in both Ca²⁺ entry across the plasma membrane and Ca²⁺ release from lysosomes⁸, and has been linked to diseases such as ischaemia-reperfusion injury, bipolar disorder and Alzheimer's disease^{9–11}. Here we report the cryo-electron microscopy structures of the zebrafish TRPM2 in the apo resting (closed) state and in the ADPR/Ca²⁺-bound active (open) state, in which the characteristic NUDT9-H domains hang underneath the MHR1/2 domain. We identify an ADPR-binding site located in the bi-lobed structure of the MHR1/2 domain. Our results provide an insight into the mechanism of activation of the TRPM channel family and define a framework for the development

of therapeutic agents to treat neurodegenerative diseases and temperature-related pathological conditions.

TRPM2 shares the characteristic N-terminal homology regions (MHR1–MHR4) and C-terminal coiled-coil domains (CTD) with the other members of the TRPM family, with its structural characteristic being its C-terminal NUDT9-H domain, a homologue of the human ADP-ribose pyrophosphatase NUDT9¹². However, it is not clear how the NUDT9-H domain is assembled in TRPM2 because of the lack of a full-length structure. In human TRPM2 (*HsTRPM2*), deletion of the NUDT9-H domain strongly decreases TRPM2 expression in the plasma membrane and abolishes channel gating, demonstrating the vital role of this domain in channel assembly, surface trafficking and channel function^{13,14}. By contrast, the TRPM2 from the starlet sea anemone *Nematostella vectensis* (*NvTRPM2*) is gated independently of the NUDT9-H domain, indicating different TRPM2 gating mechanisms between vertebrates and invertebrates¹⁵. In the presence of calcium, TRPM2 is activated by ADPR, a secondary messenger that is

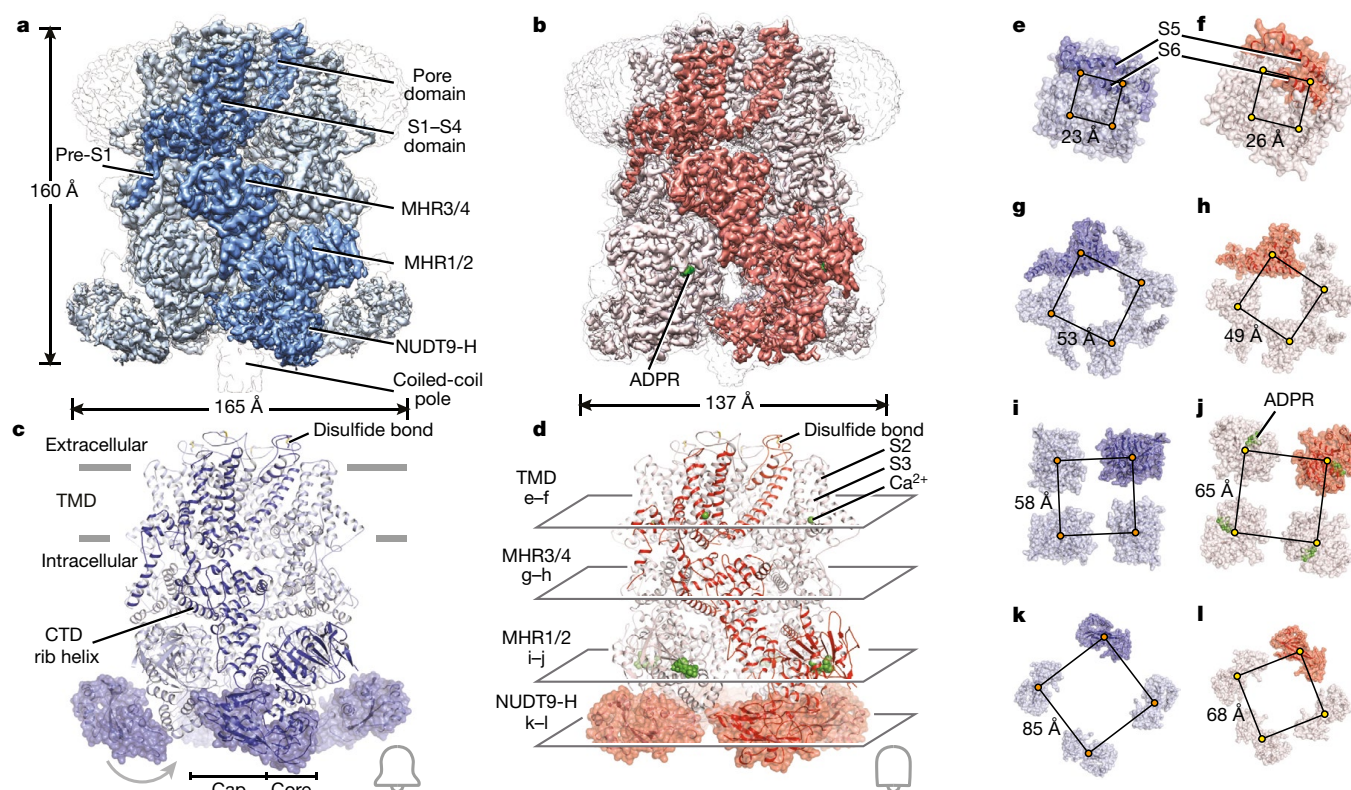


Fig. 1 | The overall architecture. **a, b**, The three-dimensional reconstruction of EDTA-TRPM2 (**a**, blue) and ADPR/Ca²⁺-TRPM2 (**b**, red) viewed parallel to the membrane. The unsharpened reconstructions are shown as transparent envelopes. One subunit is highlighted. **c, d**, Structures of the corresponding reconstructions in **a** and **b**. **e–l**, Structures

of the corresponding reconstructions shown in **a** and **b**, viewed from the intracellular side of the membrane representing the subunit arrangement in the pore domain (**e, f**), MHR3 and MHR4 domains (**g, h**), MHR1/2 domain (**i, j**) and NUDT9-H domain (**k, l**). The distances between the centre of mass of each subunit in each layer are indicated.

¹Van Andel Research Institute, Grand Rapids, MI, USA. ²Vollum Institute, Oregon Health & Science University, Portland, OR, USA. ³Present address: Janelia Research Campus, Ashburn, VA, USA. *e-mail: wei.lu@vai.org; juan.du@vai.org

released upon oxidative stress and from many metabolic processes^{12–14}. Functional studies have provided a consensus view that ADPR binds to the NUDT9-H domain^{16–18} but proof of the ADPR-binding site is lacking and it is not known how binding triggers channel opening^{15,19,20}. To understand the molecular basis and mechanisms of the channel activation of TRPM2, we determined cryo-electron microscopy structures of the full-length zebrafish (*Danio rerio*) TRPM2 (*DrTRPM2*) in the presence of EDTA and ADPR with Ca^{2+} (termed EDTA-TRPM2 and ADPR/ Ca^{2+} -TRPM2, respectively) at estimated overall resolutions of 3.8 Å and 3.3 Å, respectively (Fig. 1a–d, Extended Data Figs. 1–4, Extended Data Table 1).

EDTA-TRPM2 adopts an overall bell-like shape²¹, with the transmembrane domain as the shoulder of the bell, MHR1–MHR4 as the waist, and the expanded NUDT9-H domains as the lip (Fig. 1a, c, Extended Data Fig. 4a–c). The pore loop (P loop) is connected to the pore-lining helix S6 through a long loop in which Cys1012 and Cys1024 form a disulfide bond, stabilizing the integrity of the pore region. This is consistent with the finding that mutations to Cys1012 and/or Cys1024 markedly reduce the function of the channel²² (Fig. 1a, c, Extended Data Fig. 4a–c). In comparison to the structure of EDTA-TRPM2, ADPR/ Ca^{2+} -TRPM2 shows a distinct overall architecture (Fig. 1b, d, Extended Data Fig. 4a, d, e). The most obvious difference is a contracted NUDT9-H layer (Fig. 1a–d, k–l). We did not observe any obvious conformational change of NUDT9-H upon the addition of ADPR and, owing to limited resolution of this domain, we were unable to tell whether it binds ADPR (Extended Data Figs. 1, 3, 5). However, we identified an unambiguous density for the ADPR molecule at the cleft of the bi-lobed clamshell-like MHR1/2 (Fig. 1b, d, j, Extended Data Fig. 4d, e). The MHR1/2 layer showed expansion upon binding of agonists, with rearrangement and conformational changes in each subunit (Fig. 1i, j). Notably, both of our *DrTRPM2* structures are overall very different from the recently published structure of *NvTRPM2*²³. Nevertheless, the bi-lobed MHR1/2 in EDTA-*DrTRPM2* has an open conformation, similar to that of *NvTRPM2* and *HsTRPM4*²⁴. By contrast, the MHR1/2 of ADPR/ Ca^{2+} -*DrTRPM2* is in a closed conformation (Extended Data Fig. 6a–f).

The MHR3/4 domains frame a square hollow in the EDTA-TRPM2 structure (Fig. 1g). Upon binding of agonists, each domain is tilted, contracting the internal wall into a smaller square with a clockwise rotation, accompanied by obvious changes at the subunit interfaces (Fig. 1h). In the presence of ADPR with Ca^{2+} , the pore domain in the transmembrane domain layer shows marked expansion and rotation with a notably enlarged ion pore (Fig. 1e, f). On the basis of the different functions of each layer, we define the bottom layer—consisting of the NUDT9-H and MHR1/2 domains—as the ligand-sensing layer, and the middle layer with the MHR3 and MHR4 domains as the linker layer in our subsequent discussion of signal transduction.

To define the functional states of the *DrTRPM2* structures and to understand why TRPM2 is non-selective, we studied the ion-conducting pores (Fig. 2a, b). Similar to other TRP channels, the ion-conducting pore of TRPM2 is restricted at two locations—one is a selectivity filter located close to the extracellular entrance, formed by a hinge connecting the pore helix (P helix) and the P loop, and the other is a gate near the intracellular end lined by S6 (Fig. 2a, b). Although the two structures share a similar configuration of the selectivity filter, they show marked differences in the shape of the pore vestibules and at the gates (Fig. 2a–c). Overall, EDTA-TRPM2 has a triangular pore, similar to that of reported TRP channel structures^{23–28} (Fig. 2a, c), whereas ADPR/ Ca^{2+} -TRPM2 has an expanded pore and an elevated gate (Fig. 2b, c).

In EDTA-TRPM2, the gate is defined by Gln1068 and has a radius of 1.0 Å, which prevents ion permeation (Fig. 2a, c, d). By contrast, ADPR/ Ca^{2+} -TRPM2 has a substantially enlarged gate of radius 2.6 Å (Fig. 2b, c, e), which is large enough to pass hydrated Ca^{2+} and Na^{+} ions. Therefore, EDTA-TRPM2 represents an apo resting, non-conducting state and ADPR/ Ca^{2+} -TRPM2 represents an agonist-bound open, conducting state. Notably, despite sharing conserved

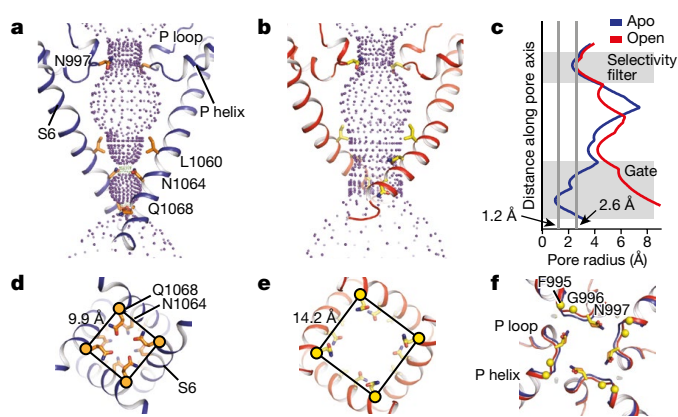


Fig. 2 | The ion-conducting pore. **a, b**, The shape and size of the ion-conducting pore in EDTA-TRPM2 (**a**) and ADPR/ Ca^{2+} -TRPM2 (**b**). Purple, green and red spheres define radii of >2.3 Å, 1.2–2.3 Å and <1.2 Å, respectively. **c**, Plot of pore radius along the pore axis. **d, e**, The gates of EDTA-TRPM2 (**d**) and ADPR/ Ca^{2+} -TRPM2 (**e**) viewed from the intracellular side. The distances between the $\text{C}\alpha$ atoms of Gln1068 in adjacent subunits are indicated. **f**, The selectivity filter in EDTA-TRPM2 (blue) and in ADPR/ Ca^{2+} -TRPM2 (red) superimpose well. The view is from the extracellular side.

residues in S6 with TRPM4, the restriction site of TRPM2 moves towards the intracellular side by one helical turn, giving rise to the outward tilting of S6 at the intracellular half (Extended Data Fig. 6j–m).

Although TRPM2 is a Ca^{2+} -permeable channel, the residues that form the selectivity filter—Phe995–Gly996–Asn997 in *DrTRPM2*—are highly conserved between TRPM2 and the Ca^{2+} -impermeable TRPM4 and TRPM5 (Fig. 2f). This sequence lacks the determinant acidic residue required for calcium permeability, as found in TRPM1, TRPM3, TRPM6 and TRPM7^{29–32}. Indeed, replacing Asn997 in *HsTRPM2* by a glutamate substantially increased calcium permeability³³. The question then is why TRPM2 is permeable to Ca^{2+} . An unusually short selectivity filter is observed in *DrTRPM2*, consisting of a single neutral residue, Asn997, which gives rise to a long and flat P helix–P loop hinge (Fig. 2a, b, f). This is distinct from the available TRP channel structures, for which the P helix–P loop hinges are shorter and thus the selectivity filters are longer and narrower^{24,26,28} (Extended Data Fig. 6g–i, l, m). We speculate that a short selectivity filter may reduce the selectivity to ions. Despite the similar configuration of the selectivity filter in *NvTRPM2* compared to that of *HsTRPM4* (Extended Data Fig. 6i), *NvTRPM2* is still Ca^{2+} -permeable because it has a glutamate instead of a glutamine residue²³. The difference in selectivity filters reveals that distinct mechanisms underlie the Ca^{2+} -permeation of TRPM2 in vertebrates and invertebrates.

From the density map of ADPR/ Ca^{2+} -TRPM2, an ADPR molecule was unambiguously identified in the cleft of MHR1/2 (Figs. 1b, d, j, 3a, b, Extended Data Figs. 3f, 4d, e). Whereas the adenine moiety stacks with Tyr271, the terminal ribose forms a hydrogen bond with the side chain of Arg278 (Fig. 3a, b). The charged alpha-phosphate moiety is in close contact with the N terminus of $\alpha 7$, and the beta-phosphate moiety is hydrogen-bonded to the side chains of Arg334 and Asn155 (Fig. 3a). Finally, the loop connecting $\beta 5$ and $\alpha 3$ lines the side of ADPR, forming interactions through the backbone with the adenine and terminal ribose moieties (Fig. 3a, b). Notably, the residues interacting with ADPR are highly conserved only among the TRPM2 orthologues (Extended Data Fig. 7). Superimposition of the MHR1/2 domains of EDTA-TRPM2 and ADPR/ Ca^{2+} -TRPM2 revealed that the binding of ADPR closes the clamshell, pulling the NUDT9-H domain towards the coiled-coil pole and thus rearranging the ligand-sensing layer (Fig. 3c–h).

To set this knowledge of the ADPR-binding site in the context of the physiological function of TRPM2, we performed electrophysiological experiments on wild-type *DrTRPM2* and its mutants in which key

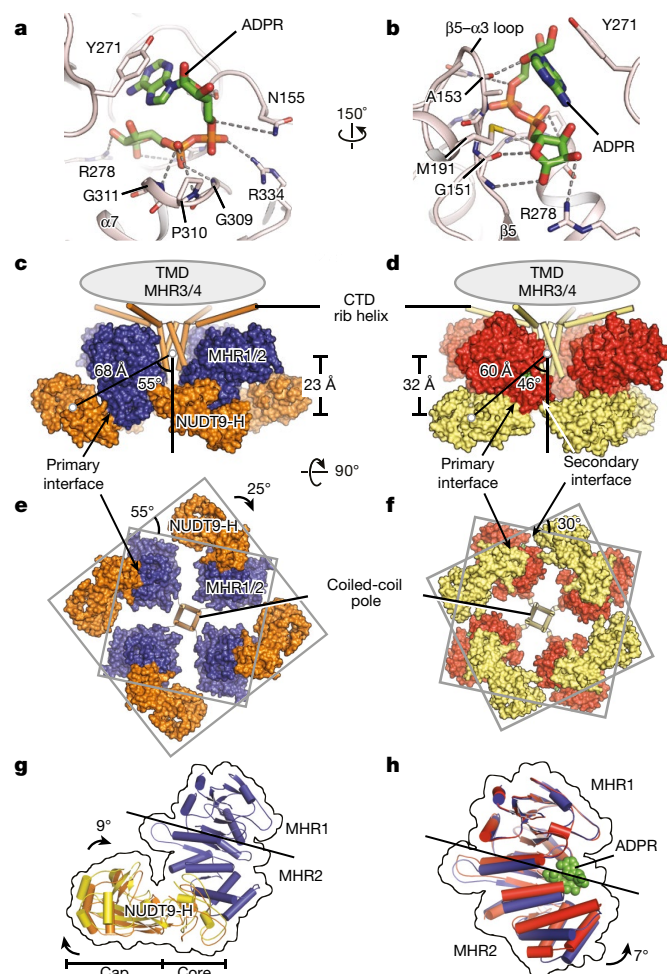


Fig. 3 | The ligand-sensing layer. **a, b**, The ADPR-binding pocket. **c–f**, Conformational changes of NUDT9-H and MHR1/2 upon ADPR binding. The distances between the centre of mass of the NUDT9-H domain and the MHR1/2 are measured in EDTA-TRPM2 (**c**) and ADPR/Ca²⁺-TRPM2 (**d**). The angles between the centre of mass of the NUDT9-H domain and the CTD coiled-coil pole along the pore axis are indicated in EDTA-TRPM2 (**e**) and ADPR/Ca²⁺-TRPM2 (**f**). **e** and **f** are viewed from the intracellular side. **g**, Comparison of the NUDT9-H domains of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2, by superimposition of the MHR2 domains. The rotation of NUDT9-H relative to MHR1/2 upon the binding of ADPR is measured. For clarity, only the MHR1/2 domain of EDTA-TRPM2 is shown. **h**, Superimposition of the MHR1 domains of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 shows closure of the clamshell of the MHR1/2 domain upon ADPR binding.

residues within the ADPR-binding site were replaced (Extended Data Fig. 8a–d). Most mutations showed a decrease in agonist-induced current, and the double mutant R278A/R334A nearly abolished the current (Extended Data Fig. 8c, d). In addition, Met191—another residue that shapes the ADPR-binding site—has been proposed as a sensitization sensor for redox signalling molecules such as H₂O₂³⁴. These results support the idea that the MHR1/2 domain acts as a physiologically relevant ligand-binding site, and also provide a plausible explanation of the mechanism underlying redox sensitization by which oxidation of this methionine perhaps mimics or enhances the binding of ADPR.

To understand how the binding of ADPR initiates channel activation and to investigate a possible role of the NUDT9-H domain, we compared the ligand-sensing layer between two *Dr*TRPM2 structures and performed electrophysiology experiments. In EDTA-TRPM2, the NUDT9-H domain is positioned at the extreme left of MHR1/2 through a primary interface, and lacks interactions with the rest of the channel (Fig. 3c, e, g). Consequently, NUDT9-H exhibits a high degree

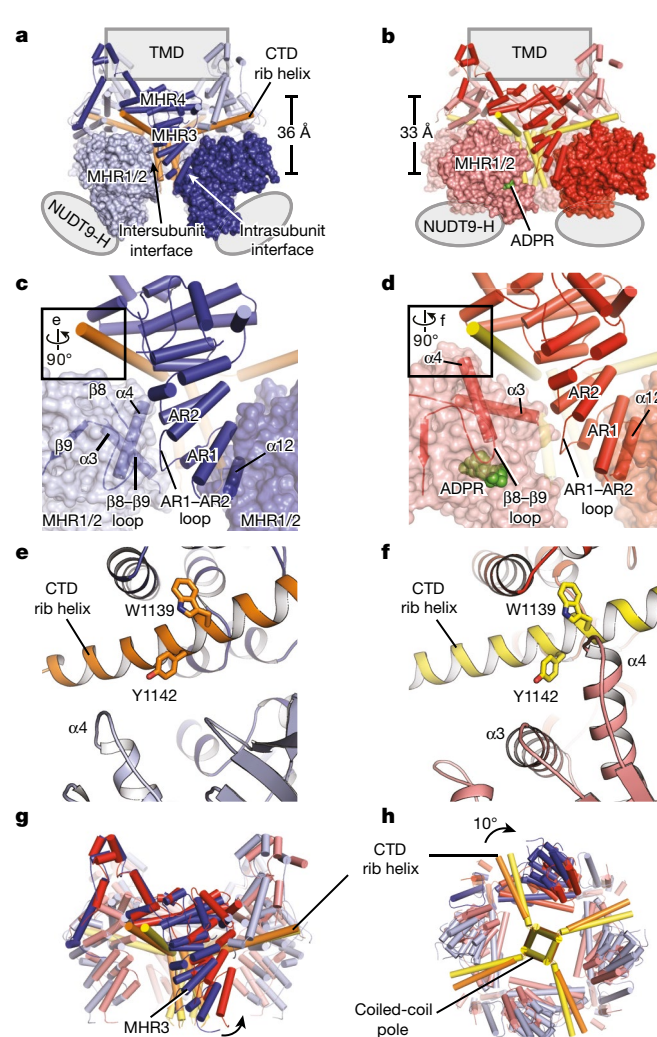


Fig. 4 | Signal transduction through the linker layer MHR3/4. **a, b**, The intercellular and intracellular interfaces of the linker layer in EDTA-TRPM2 (**a**, blue) and ADPR/Ca²⁺-TRPM2 (**b**, red). Distances between the centre of masses of MHR1/2 and MHR3/4 are measured (**b**). **c, d**, Conformational changes at the inter- and intrasubunit interactions between EDTA-TRPM2 (**c**) and ADPR/Ca²⁺-TRPM2 (**d**). **e, f**, A new interface between the CTD rib helix and MHR1 is created upon the binding of ADPR and Ca²⁺. Trp1139 and Tyr1142 interacting with α 4 in ADPR/Ca²⁺-TRPM2 are shown as sticks. **g, h**, Superimposition of the linker layers of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2, by aligning the CTD coiled-coil poles (residues Val1159 to Val1184).

of flexibility (Extended Data Fig. 4a, b). In the ADPR/Ca²⁺-TRPM2 structure, NUDT9-H is displaced towards the pore axis (Fig. 3c–f), creating a loose secondary interface with the adjacent MHR2 domain, thus stabilizing the expanded MHR1/2 layer and possibly also the activated conformation (Fig. 3d, f). Indeed, removal of NUDT9-H abolished the ADPR/Ca²⁺-induced current (Extended Data Fig. 8c, d), consistent with the idea that NUDT9-H is indispensable for channel gating in *Hs*TRPM2^{13,14}.

The MHR3/4 domain of *Dr*TRPM2 has a triangular shape, with MHR3 as the vertical edge wedged into the space between adjacent MHR1/2 domains. This results in two interfaces: the intra- and inter-subunit MHR1/2–MHR3/4 interfaces (Fig. 4a–d). An ATP molecule has been identified at the intersubunit interface in the structure of TRPM4, indicating the important role of this interface in the sensing of stimuli²⁶. Signal transduction from the MHR1/2 domain upon binding of ADPR has two major consequences. First, the intersubunit MHR1/2–MHR3/4 interface is markedly changed (Fig. 4c, d). The α 3 in the MHR1/2 shifts towards MHR3/4, and uses its N terminus to push

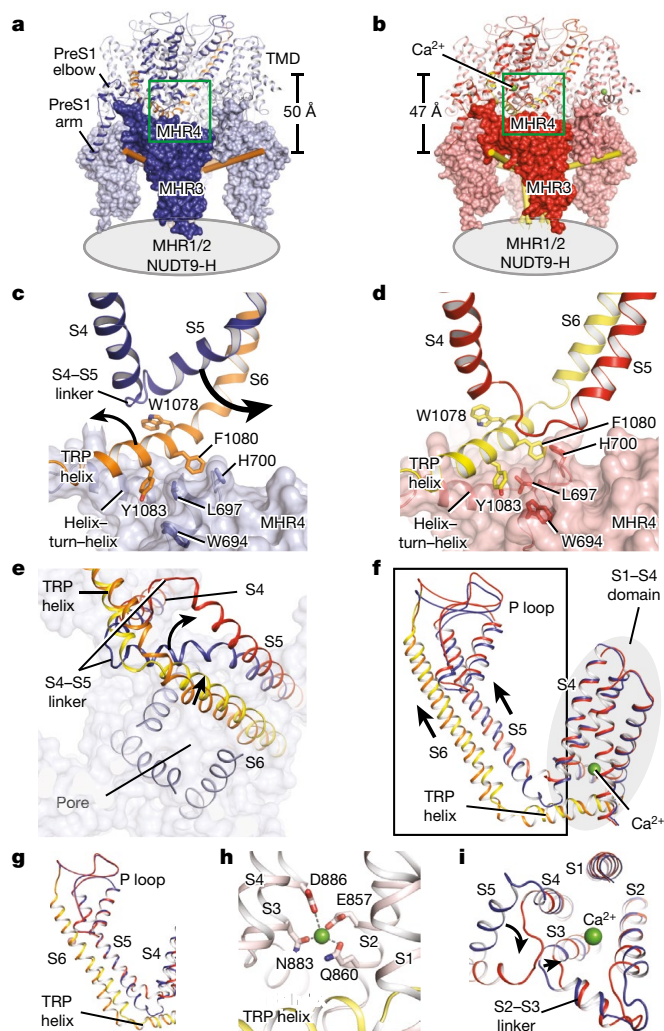


Fig. 5 | Signal transduction in the transmembrane domain from EDTA-TRPM2 to ADPR/Ca²⁺-TRPM2. **a, b**, Interface between the linker layer and the transmembrane domain layer. Distances between the centre of masses of MHR3/4 and the transmembrane domain are measured. **c, d**, A dovetailed interaction between MHR4 and the TRP helix. The movements of the S4-S5 linker and the TRP helix are indicated. **e, f**, Superimposition of the EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 structures of *Dr*TRPM2, by aligning the transmembrane domain layers (e) and the S1-S4 domains (f), viewed from the intracellular side. One single subunit is highlighted. **g**, Superimposition of the EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 structures by aligning the pore domains, viewed parallel to the membrane. **h**, The calcium-binding sites of TRPM2. **i**, Superimposition of the EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 structures by aligning the S1-S4 domains highlights the conformational changes of S3 and the S2-S3 linker that are induced by calcium binding.

MHR3/4 towards the upper right (Fig. 4c, d). Second, ADPR binding promotes the movement of $\alpha 4$ in the MHR1/2 towards the CTD rib helix (Fig. 4e, f)—creating a new interface between the CTD rib helix and MHR1/2—and the C terminus of $\alpha 4$ is clamped between two bulky residues in the rib helix, producing an anticlockwise rotation of the rib helix around the pore axis (Fig. 4g, h). Taken together, these findings suggest that MHR3/4 in the linker layer has a key role in conveying signals that initiate from the ligand-sensing layer to the transmembrane domain, and acts by modification of their interfaces.

The communication between the intracellular domain and the transmembrane domain is mainly mediated by the TRP helix³⁵. In EDTA-TRPM2, the TRP helix tightly interacts with the S4-S5 linker on the top through a highly conserved Trp1078 residue on the TRP helix, and with the MHR4 on the bottom (Fig. 5a–d)—thus fixing S5 in a flexed

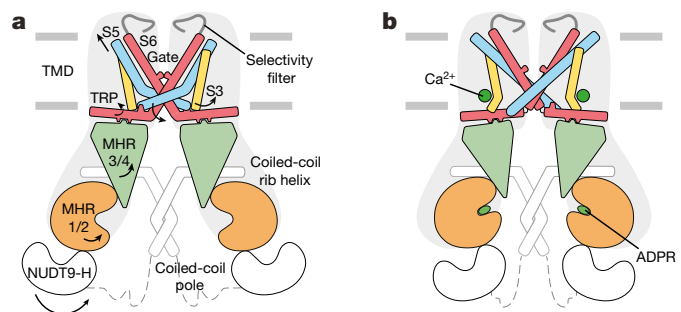


Fig. 6 | Schematic of the activation mechanism of TRPM2.

a, b, Conformational differences between EDTA-TRPM2 (a) and ADPR/Ca²⁺-TRPM2 (b). The major movements induced upon ADPR/Ca²⁺ binding are shown by arrows.

conformation (Fig. 5c–e). Upon binding of ADPR, the upward motion of MHR4 is conveyed to the TRP helix by pushing it towards the S4-S5 linker (Fig. 5c–e). As a result, the restriction of the S4-S5 linker by Trp1078 is released. The intracellular end of S5 swaps from one side of the TRP helix to the other, thus adopting a relatively straightened conformation, and moves up towards the extracellular side by one half helical turn (Fig. 5c–f). These structural rearrangements of S5 not only promote the translation of S6 upward by approximately one half helical turn but also free a space at the intracellular end. This enables an outward tilting of the intracellular region of S6, where the gate is located, thus opening the channel (Fig. 5c–f). Notably, we observed a putative calcium density located close to S3, coordinated by residues which are highly conserved in Ca²⁺-dependent members of the TRPM family (Fig. 5f, h, i, Extended Data Fig. 9). Comparison of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 shows the repositioning of S3 upon the binding of calcium. We propose that this frees up space for the relocation of the S4-S5 linker and thus facilitates channel activation. Mutation studies of Asn883 and Asp886 in the binding pocket have been reported to affect the function of the TRPM2 channel³⁶. Superimposition of the pore domains of the two TRPM2 structures revealed that conformational changes are restricted to the intracellular region (Fig. 5g), emphasizing the interplay between the TRP helix and the S4-S5 linker as the key determinant of the activation of TRPM channels.

An agonist-induced activation mechanism is apparent from the analysis of two *Dr*TRPM2 structures, in which notable motion transduction is seen along three layers from the ligand-sensing layer, along the linker layer, to the very distal transmembrane domain (Fig. 6). Binding of ADPR induces the closure of bi-lobed MHR1/2 in the ligand-sensing layer, accompanied by the swing of NUDT9-H towards the pore axis, which creates an interface between NUDT9-H and the adjacent MHR1/2. The motion is further translated to the linker layer, with MHR3/4 moving towards the upper right, leading to a repositioning of the TRP helix. This unlocks the restriction of the S4-S5 linker and enables S5 to relocate from one side of the TRP helix to the other, changing from a flexed to a straightened conformation. We speculate that calcium facilitates the relocation of the S4-S5 linker by tilting the nearby S3. Finally, the relocation of S5 causes an outward tilting of S6, which results in opening of the channel. The NUDT9-H domain has a crucial role in channel gating, perhaps by stabilizing the channel in an open conformation. Whether it has a second binding site for ADPR, or acts as a stimuli sensory domain to modulate the channel functions, is yet to be determined.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0558-4>.

Received: 12 April 2018; Accepted: 3 August 2018;

Published online 24 September 2018.

1. Tan, C. H. & McNaughton, P. A. The TRPM2 ion channel is required for sensitivity to warmth. *Nature* **536**, 460–463 (2016).
2. Song, K. et al. The TRPM2 channel is a hypothalamic heat sensor that limits fever and can drive hypothermia. *Science* **353**, 1393–1398 (2016).
3. Knowles, H. et al. Transient Receptor Potential Melastatin 2 (TRPM2) ion channel is required for innate immunity against *Listeria monocytogenes*. *Proc. Natl Acad. Sci. USA* **108**, 11578–11583 (2011).
4. Hecquet, C. M. et al. Cooperative interaction of *trp* melastatin channel transient receptor potential (TRPM2) with its splice variant TRPM2 short variant is essential for endothelial cell apoptosis. *Circ. Res.* **114**, 469–479 (2014).
5. Kolisek, M., Beck, A., Fleig, A. & Penner, R. Cyclic ADP-ribose and hydrogen peroxide synergize with ADP-ribose in the activation of TRPM2 channels. *Mol. Cell* **18**, 61–69 (2005).
6. Beck, A., Kolisek, M., Bagley, L. A., Fleig, A. & Penner, R. Nicotinic acid adenine dinucleotide phosphate and cyclic ADP-ribose regulate TRPM2 channels in T lymphocytes. *FASEB J.* **20**, 962–964 (2006).
7. Togashi, K. et al. TRPM2 activation by cyclic ADP-ribose at body temperature is involved in insulin secretion. *EMBO J.* **25**, 1804–1815 (2006).
8. Lange, I. et al. TRPM2 functions as a lysosomal Ca^{2+} -release channel in β cells. *Sci. Signal.* **2**, ra23 (2009).
9. Miller, B. A. et al. TRPM2 channels protect against cardiac ischemia-reperfusion injury: role of mitochondria. *J. Biol. Chem.* **289**, 7615–7629 (2014).
10. Xu, C. et al. Association of the putative susceptibility gene, transient receptor potential protein melastatin type 2, with bipolar disorder. *Am. J. Med. Genet. B* **141B**, 36–43 (2006).
11. Ostapchenko, V. G. et al. The transient receptor potential melastatin 2 (TRPM2) channel contributes to β -amyloid oligomer-related neurotoxicity and memory impairment. *J. Neurosci.* **35**, 15157–15169 (2015).
12. Perraud, A. L. et al. ADP-ribose gating of the calcium-permeable LTRPC2 channel revealed by Nudix motif homology. *Nature* **411**, 595–599 (2001).
13. Perraud, A. L. et al. Accumulation of free ADP-ribose from mitochondria mediates oxidative stress-induced gating of TRPM2 cation channels. *J. Biol. Chem.* **280**, 6138–6148 (2005).
14. Wehage, E. et al. Activation of the cation channel long transient receptor potential channel 2 (LTRPC2) by hydrogen peroxide. *J. Biol. Chem.* **277**, 23150–23156 (2002).
15. Kühn, F. J., Kühn, C., Winking, M., Hoffmann, D. C. & Lückhoff, A. ADP-ribose activates the TRPM2 channel from the sea anemone *Nematostella vectensis* independently of the NUDT9H domain. *PLoS ONE* **11**, e0158060 (2016).
16. Kühn, F. J. & Lückhoff, A. Sites of the NUDT9-H domain critical for ADP-ribose activation of the cation channel TRPM2. *J. Biol. Chem.* **279**, 46431–46437 (2004).
17. Yu, P. et al. Identification of the ADPR binding pocket in the NUDT9 homology domain of TRPM2. *J. Gen. Physiol.* **149**, 219–235 (2017).
18. Fliegert, R. et al. Ligand-induced activation of human TRPM2 requires the terminal ribose of ADPR and involves Arg1433 and Tyr1349. *Biochem. J.* **474**, 2159–2175 (2017).
19. Iordanov, I., Mihályi, C., Tóth, B. & Csányi, L. The proposed channel-enzyme transient receptor potential melastatin 2 does not possess ADP ribose hydrolase activity. *eLife* **5**, e17600 (2016).
20. Tóth, B., Iordanov, I. & Csányi, L. Putative chanzyme activity of TRPM2 cation channel is unrelated to pore gating. *Proc. Natl Acad. Sci. USA* **111**, 16949–16954 (2014).
21. Maruyama, Y. et al. Three-dimensional reconstruction using transmission electron microscopy reveals a swollen, bell-shaped structure of transient receptor potential melastatin type 2 cation channel. *J. Biol. Chem.* **282**, 36961–36970 (2007).
22. Mei, Z. Z., Mao, H. J. & Jiang, L. H. Conserved cysteine residues in the pore region are obligatory for human TRPM2 channel function. *Am. J. Physiol. Cell Physiol.* **291**, C1022–C1028 (2006).
23. Zhang, Z., Toth, B., Szollosi, A., Chen, J. & Csányi, L. Structure of a TRPM2 channel in complex with Ca^{2+} explains unique gating regulation. *eLife* **7**, (2018).
24. Winkler, P. A., Huang, Y., Sun, W., Du, J. & Lü, W. Electron cryo-microscopy structure of a human TRPM4 channel. *Nature* **552**, 200–204 (2017).
25. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
26. Guo, J. et al. Structures of the calcium-activated, non-selective cation channel TRPM4. *Nature* **552**, 205–209 (2017).
27. Yin, Y. et al. Structure of the cold- and menthol-sensing ion channel TRPM8. *Science* **359**, 237–241 (2018).
28. Autzen, H. E. et al. Structure of the human TRPM4 ion channel in a lipid nanodisc. *Science* **359**, 228–232 (2018).
29. Grimm, C., Kraft, R., Sauerbruch, S., Schultz, G. & Harteneck, C. Molecular and functional characterization of the melastatin-related cation channel TRPM3. *J. Biol. Chem.* **278**, 21493–21501 (2003).
30. Nadler, M. J. et al. LTRPC7 is a Mg-ATP-regulated divalent cation channel required for cell viability. *Nature* **411**, 590–595 (2001).
31. Voets, T. et al. TRPM6 forms the Mg^{2+} influx channel involved in intestinal and renal Mg^{2+} absorption. *J. Biol. Chem.* **279**, 19–25 (2004).
32. Lambert, S. et al. Transient receptor potential melastatin 1 (TRPM1) is an ion-conducting plasma membrane channel inhibited by zinc ions. *J. Biol. Chem.* **286**, 12221–12233 (2011).
33. Xia, R. et al. Identification of pore residues engaged in determining divalent cationic permeation in transient receptor potential melastatin subtype channel 2. *J. Biol. Chem.* **283**, 27426–27432 (2008).
34. Kashio, M. et al. Redox signal-mediated sensitization of transient receptor potential melastatin 2 (TRPM2) to temperature affects macrophage functions. *Proc. Natl Acad. Sci. USA* **109**, 6745–6750 (2012).
35. Gregorio-Teruel, L. et al. The integrity of the TRP domain is pivotal for correct TRPV1 channel gating. *Biophys. J.* **109**, 529–541 (2015).
36. Winking, M. et al. Importance of a conserved sequence motif in transmembrane segment S3 for the gating of human TRPM8 and TRPM2. *PLoS ONE* **7**, e49877 (2012).

Acknowledgements We thank G. Zhao and X. Meng for support with data collection at the David Van Andel Advanced Cryo-Electron Microscopy Suite, the HPC team in the Van Andel Research Institute (VARI) for computational support, C. Xu for help with SerialEM, and D. Nadziejka for technical editing. This work was supported by internal VARI funding.

Reviewer information Nature thanks A. H. Guse, A. Ward and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.D. and W.L. initiated the project. Y.H. and P.A.W. purified TRPM2; Y.H. and W.S. performed electrophysiological experiments; and Y.H., J.D. and W.L. performed cryo-electron microscopy data collection, processing, analysis and wrote the manuscript. All authors contributed to the preparation of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0558-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0558-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to W.L. or J.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

TRPM2 construct design. The human, bovine, chicken, mouse and zebrafish TRPM2 genes (UniProtKB (<http://www.uniprot.org>) accession numbers O94759, E1BFK7, F1NGK6, Q91YD4 and A0A0R4IN04, respectively) were synthesized by GenScript and sub-cloned into the pEG BacMam vector containing a twin StrepII tag, an octa-histidine tag, and enhanced green fluorescent protein (eGFP) with thrombin cleavage site (Leu-Val-Pro-Arg-Gly-Ser) at the N terminus of the gene for purification³⁷. Fluorescence-detection size-exclusion chromatography was used for initial construct screening³⁸. We screened several TRPM2 homologues on the basis of their biochemical stability and identified the TRPM2 from *D. rerio*, which has 65% overall sequence similarity to its human homologue, as a promising candidate for structural studies (Extended Data Fig. 7).

Expression and purification. The full-length DrTRPM2 construct was transformed into DH10Bac cells for bacmid production. The bacmid containing the TRPM2 gene was used to transfect Sf9 cells using Cellfectin II (Thermo Fisher Scientific) for baculovirus production. P2 virus was used to infect a suspension of HEK293 GnTI⁻ cells incubated at 37 °C. Sodium butyrate (10 mM) was added to the suspension 8–12 h post-infection and the temperature was adjusted to 30 °C to increase protein expression. Seventy-two hours after infection, the mammalian cells were collected and washed with TBS buffer (150 mM NaCl, 20 mM Tris HCl pH 8.0). After collection, cells were lysed in 10 mM Tris-HCl pH 8.0 buffer containing 1 mM phenylmethylsulfonyl fluoride (PMSF), 2 mM pepstatin, 0.8 μ M aprotinin and 2 μ g ml⁻¹ leupeptin for 1 h. After lysis, Tris-HCl pH 8.0 and NaCl were added to give a final concentration of 20 mM and 150 mM respectively and incubated for an additional 20 min. Cell debris was removed by centrifugation at 3,000g for 10 min. Membranes were collected by ultracentrifugation using a 45 Ti rotor at 186,000g for 20 min at 4 °C (Beckman Coulter). The membrane was then homogenized using a Dounce homogenizer in TBS buffer containing 1 mM PMSF, 2 mM pepstatin, 0.8 μ M aprotinin, 2 μ g ml⁻¹ leupeptin and 2 mM 2-mercaptoethanol. The membrane was solubilized using 10 mM glyco-diosgenin for 1 h at 4 °C before ultracentrifugation for 30 min at 186,000g. The supernatant was incubated with Talon resin (Clontech) for 2 h and then washed with six bed volumes of TBS buffer containing 0.2 mM glyco-diosgenin, 10 mM imidazole and 2 mM 2-mercaptoethanol. The protein was eluted with TBS buffer containing 0.2 mM glyco-diosgenin, 250 mM imidazole and 2 mM 2-mercaptoethanol. The protein peak fractions were concentrated and loaded onto a Superose 6 column (GE Healthcare) using TBS buffer containing 0.2 mM glyco-diosgenin and 5 mM 2-mercaptoethanol. The peak fractions were combined and concentrated to 4.5 mg ml⁻¹ using a 100-kDa concentrator (Millipore).

Electron microscopy sample preparation and data acquisition. The purified TRPM2 protein was incubated with either 1 mM EDTA or 1 mM ADPR and 1 mM CaCl₂ before grid preparation. Quantifoil holey carbon grids (Au 1.2/1.3 μ m size/hole space, 300 mesh) were glow-discharged for 30 s. Then 2.5 μ l of protein sample was added to the carbon face of the grids and blotted for 2 s with a 5-s waiting time. The grid was plunge-frozen in liquid ethane cooled by liquid nitrogen using a Vitrobot Mark III held at 18 °C and 100% humidity.

Images were taken using an FEI Titan Krios electron microscope operating at 300 kV with a nominal magnification of 130,000. Images were recorded by a Gatan K2 Summit direct electron detector operated in super-resolution counting mode with a binned pixel size of 1.088 Å. Each image was dose-fractionated to 40 frames with a total exposure time of 8 s with 0.2 s per frame. The dose rate was 6.76 e⁻ Å⁻² s⁻¹. The images were recorded using the automated acquisition program SerialEM³⁹. Nominal defocus values varied from -1.0 to -2.5 μ m.

Electron microscopy data processing. Images were motion-corrected, summed and 2 × 2 binned in Fourier space using MotionCor2⁴⁰. Defocus values were estimated using Gctf⁴¹. Particles were picked using Gautomatch (<http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/>) and subjected to an initial reference-free 2D classification using RELION⁴². Nine representative 2D class averages were selected as templates for automated particle-picking for the entire dataset using Gautomatch. The auto-picked particles were visually checked and false positives were removed. The particles were further cleaned up by several rounds of 2D classification using RELION⁴². The values of the contrast transfer function of individual particles from selected 2D class averages were estimated using Gctf⁴¹. The initial reconstruction was obtained using cryoSPARC⁴³. The particles were then classified into six classes using 3D classification function in RELION, with the initial reconstruction low-pass-filtered to 50 Å as a reference model. Particles from classes showing high-resolution features were combined and refined with C4 symmetry using RELION. Particles were further refined using the local refinement from Frealign with C4 symmetry applied and high-resolution limit for particle alignment set to 5.0 or 5.5 Å for EDTA-TRPM2 or ADPR/Ca²⁺-TRPM2, respectively⁴⁴. The resolutions reported are based on the 'limiting resolution' procedure, in which the resolution during refinement is limited to a lower resolution than the resolution estimated for the final reconstruction. The final resolutions reported in Extended Data Table 1 are based on the gold standard Fourier shell correlation

0.143 criteria. To calculate the Fourier shell correlation plot, a soft mask (5.4 Å extended from the reconstruction with an additional 5.4 Å cosine soft edge, low-pass-filtered to 10 Å) was applied to the two half maps. Local resolutions were estimated using Bsoft⁴⁵.

Model building and structural determination. The ADPR/Ca²⁺-TRPM2 model was built in Coot using the TRPM4 structure as a guide (RCSB Protein Data Bank (PDB) ID: 5WP6)^{24,46}. In EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2, the densities of the transmembrane domain—including the pore domain and the S1–S4 domain—and the MHR domain are mostly well-defined (Fig. 1a, b, Extended Data Fig. 3). By contrast, the NUDT9-H domain, which consists of a cap and a core region based on the nomenclature in the NUDT9, is less well-defined, particularly at the core region (Fig. 1a, b, Extended Data Fig. 3). For the NUDT9-H domain, a homology model was generated with the crystal structure of the human NUDT9 (PDB ID: 1Q33) as a template using the SWISS-MODEL online server⁴⁷. This homology model was rigid-body-fitted into the map using Chimera, followed by manual adjustment in Coot⁴⁸. The initial model was then subjected to molecular dynamics flexible fitting⁴⁹, followed by real space refinement using phenix.real_space_refine with secondary-structure restraints⁵⁰. The refined model was further manually examined and adjusted in Coot. The U-shaped density at the ADPR-binding site in the cleft of MHR1/2 is unambiguous for an ADPR molecule. For validation of the refined structure, Fourier shell correlation curves were applied to calculate the difference between the final model and electron microscopy map. The geometries of the atomic models were evaluated using MolProbity⁵¹ and EMRinger⁵². The EDTA-TRPM2 model was built using ADPR/Ca²⁺-TRPM2 as a guide. The NUDT9-H domain of ADPR/Ca²⁺-TRPM2 was rigid-body-fitted into the EDTA-TRPM2 map. This initial model was then subjected to the same refinement procedure as the ADPR/Ca²⁺-TRPM2 structure. The densities of the terminal half of the CTD coiled-coil pole and the connection between the CTD coiled-coil pole and NUDT9-H in both structures are not visible, indicating flexible movement of the NUDT9-H domain. All figures were prepared using UCSF Chimera⁵³ and PyMOL (<https://pymol.org>).

Electrophysiology. HEK293 cells were transfected using Lipofectamine 2000 (Thermo Fisher) according to the manufacturer's protocol. Transfected cells were incubated in a 24-well plate at 37 °C and were recorded 12–24 h post transfection. Inside-out patches were pulled from transfected cells and recordings were performed using a HEKA EPC-10 amplifier at room temperature. The holding potential was +60 mV. The electrodes were filled with internal solution containing 150 mM NaCl, 3 mM KCl and 10 mM HEPES (pH 7.4 adjusted with NaOH), and the bath solution was the same as the internal solution. A bath solution with 0.1 mM ADPR and 1 mM CaCl₂ was used for channel activation. The solution change was performed using a two-barrel theta-glass pipette controlled manually. Data were acquired at 10 kHz using Patchmaster software (HEKA). Data were filtered at 1 kHz, and analysed with Axograph software (<http://www.axograph.com>).

Data reporting. The sample sizes were based on the consistency of the recordings. For electrophysiological experiments, cells with GFP fluorescence were randomly selected. The investigators were blind to group allocation during data collection and analysis.

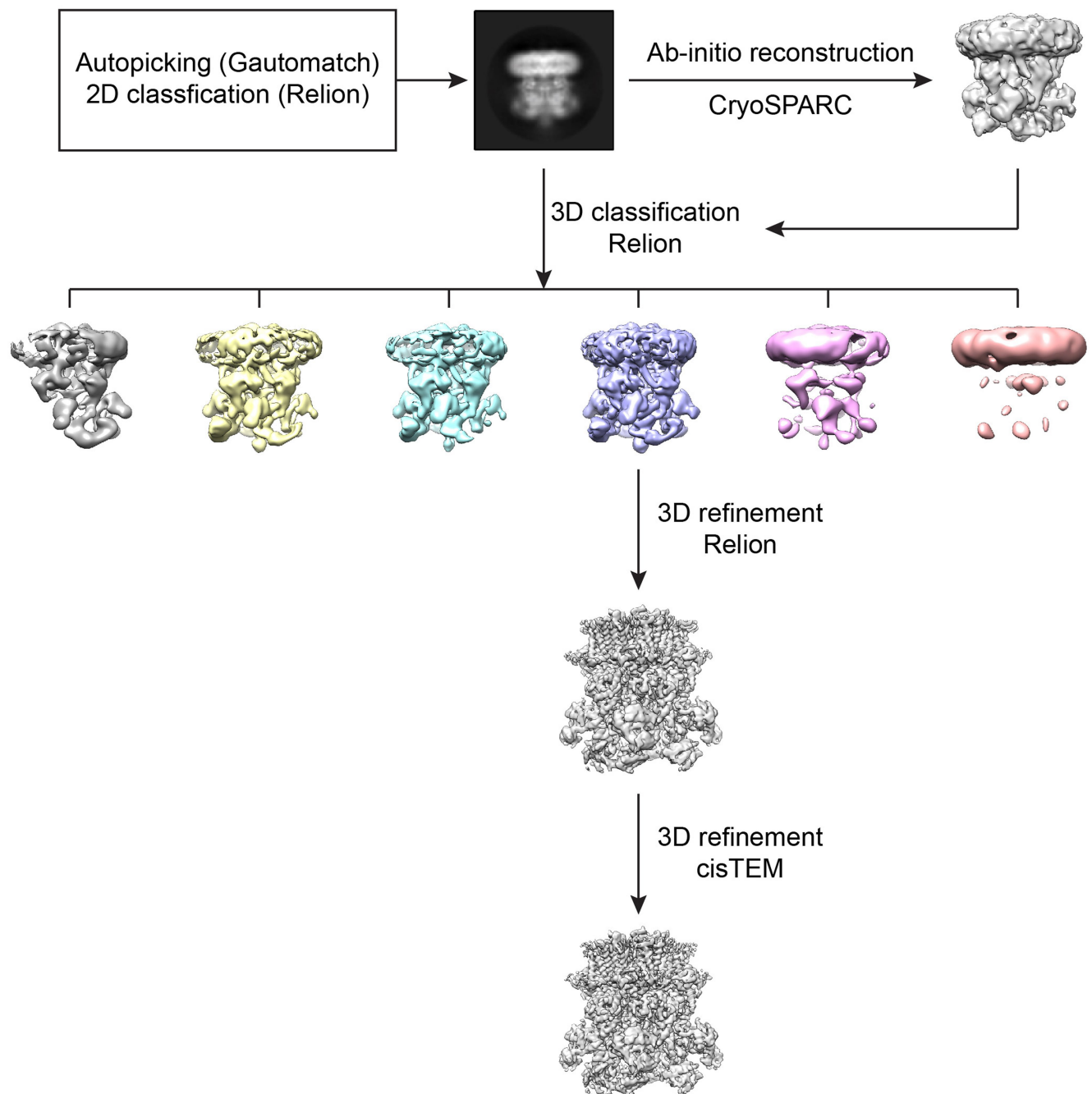
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-electron microscopy density map and coordinates of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMD-8901 and EMD-7999 and in the Research Collaboratory for Structural Bioinformatics Protein Data Bank under accession codes 6DRK and 6DRJ. All other data relating to this study are available from the corresponding author upon reasonable request.

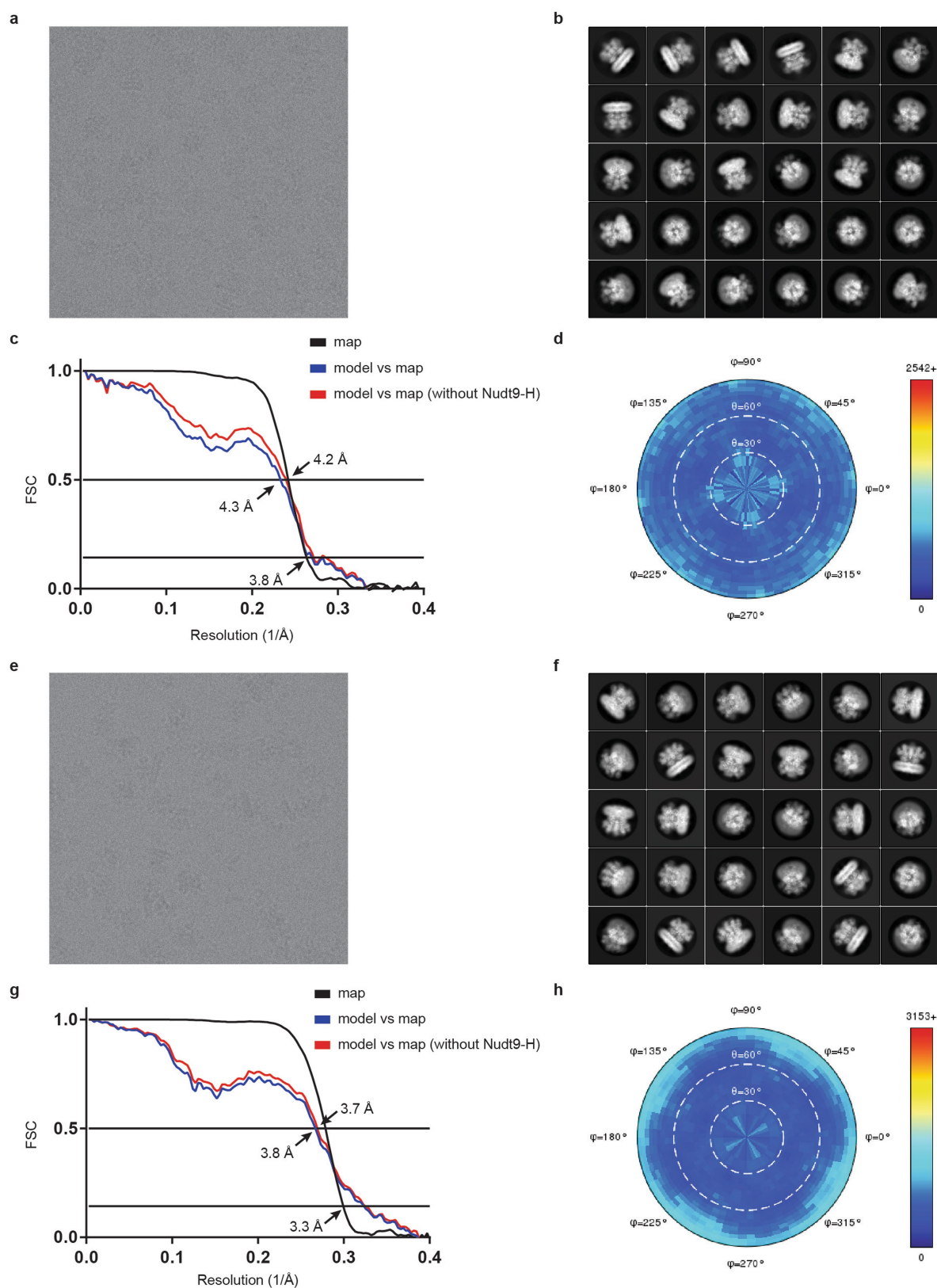
- Goehring, A. et al. Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat. Protoc.* **9**, 2574–2585 (2014).
- Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).
- Mastrorade, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
- Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
- Grigorieff, N. Frealign: an exploratory tool for single-particle cryo-EM. *Methods Enzymol.* **579**, 191–226 (2016).

45. Heymann, J. B. Guidelines for using Bsoft for high resolution reconstruction and validation of biomolecular structures from electron micrographs. *Protein Sci.* **27**, 159–171 (2018).
46. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
47. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
48. Shen, B. W., Perraud, A. L., Scharenberg, A. & Stoddard, B. L. The crystal structure and mutational analysis of human NUDT9. *J. Mol. Biol.* **332**, 385–398 (2003).
49. Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174–180 (2009).
50. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
51. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
52. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
53. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
54. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018).
55. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
56. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).



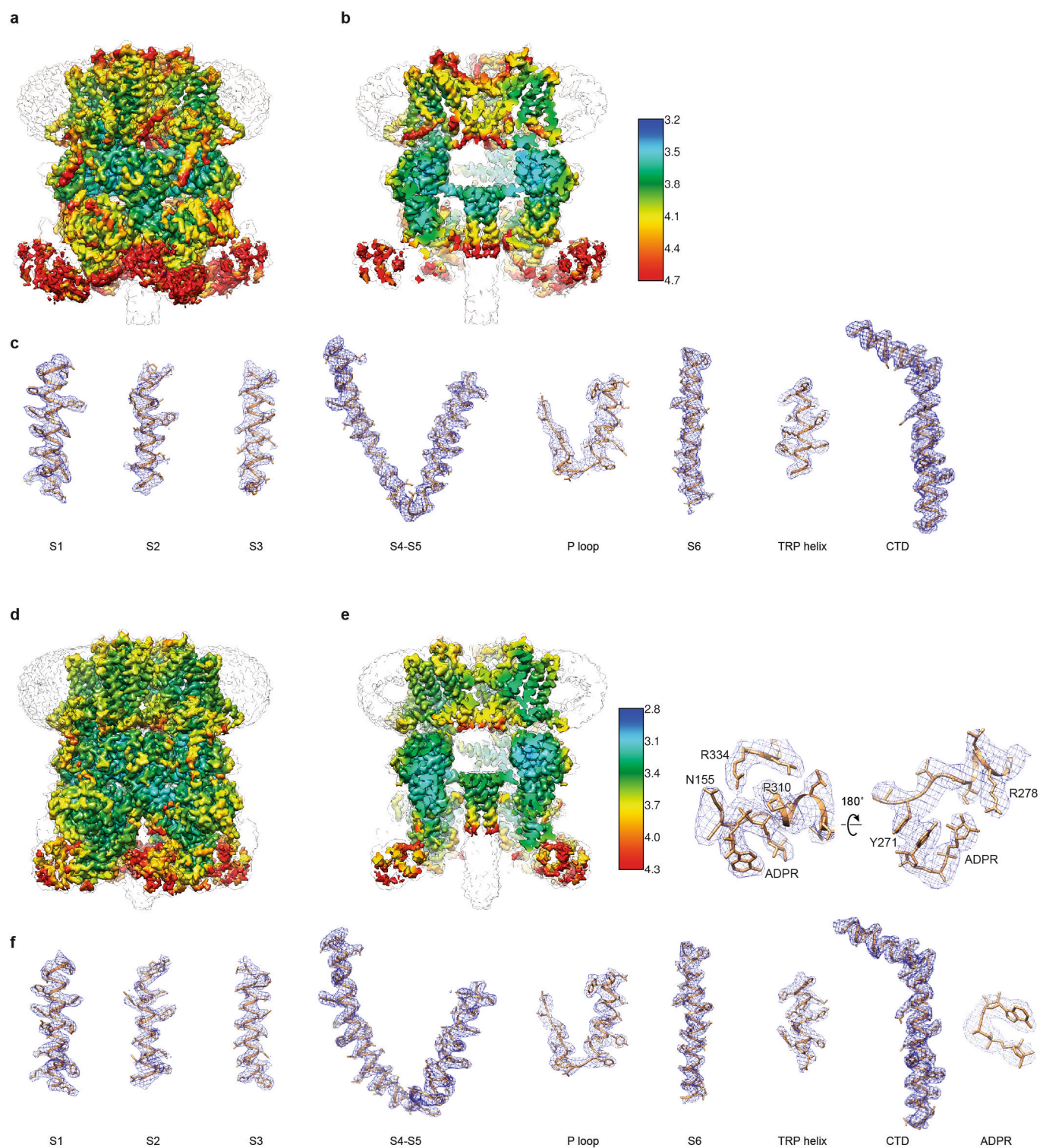
Extended Data Fig. 1 | The cryo-electron microscopy data processing workflow, using the data of EDTA-TRPM2 as an example. Particles were auto-picked using Gautomatch and visually checked in RELION. After removing false positives, particles were subjected to two rounds of 2D classification in RELION. The contrast transfer function values of

individual particles from selected 2D class averages were estimated using Gctf. For 3D classification in RELION, a reference model was generated using CryoSPARC. Initial 3D refinement was carried out using RELION. Particles were further refined using cisTEM⁵⁴.



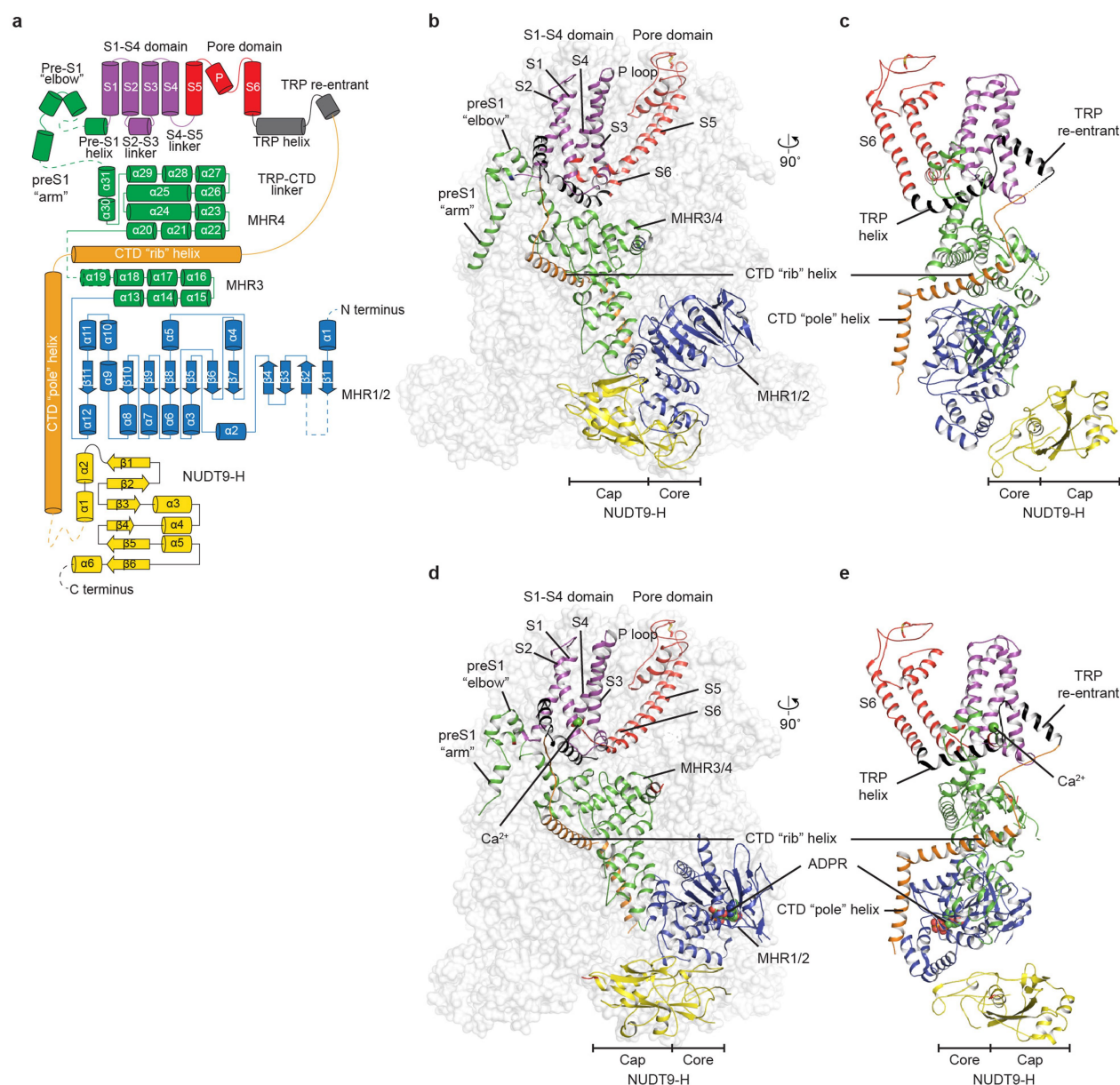
Extended Data Fig. 2 | Cryo-electron microscopy analysis of full-length *DrTRPM2* in the presence of EDTA and ADPR/Ca²⁺. **a, e**, Representative electron micrograph of EDTA-TRPM2 (a) and ADPR/Ca²⁺-TRPM2 (e). **b, f**, Selected 2D class averages of the electron micrographs of EDTA-TRPM2 (b) and ADPR/Ca²⁺-TRPM2 (f). **c, g**, The gold-standard Fourier shell correlation curves for the electron microscopy

maps of EDTA-TRPM2 (c) and ADPR/Ca²⁺-TRPM2 (g) are shown in black, and the Fourier shell correlation curves between the atomic model and the final electron microscopy maps are shown in blue. **d, h**, Angular distribution of particles used for the refinement of EDTA-TRPM2 (d) and ADPR/Ca²⁺-TRPM2 (h).



Extended Data Fig. 3 | Representative densities of the reconstruction of EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2. a, b, Local resolution estimation of the structure of EDTA-TRPM2. The map is coloured according to local resolution estimation. The unsharpened reconstructions

are shown as transparent envelopes. **c,** Representative densities of EDTA-TRPM2. **d, e,** Local resolution estimation of the structure of ADPR/Ca²⁺-TRPM2. **f,** Representative densities of ADPR/Ca²⁺-TRPM2.



Extended Data Fig. 4 | Overall architecture of TRPM2. **a**, Domain organization of TRPM2. Dashed lines and cylinders denote regions that have not been modelled. **b, c**, Cartoon representation of one subunit of the

EDTA-TRPM2 structure. **d, e**, Cartoon representation of one subunit of the ADPR/Ca²⁺-TRPM2 structure. ADPR and Ca²⁺ are shown as spheres.

a

```

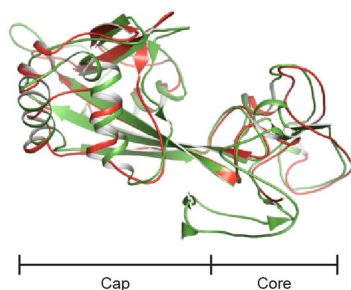
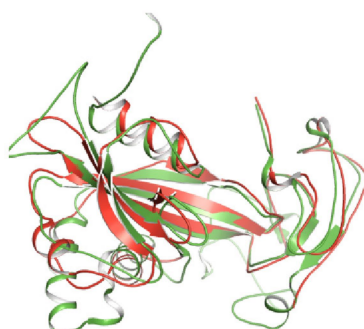
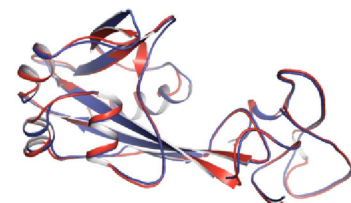
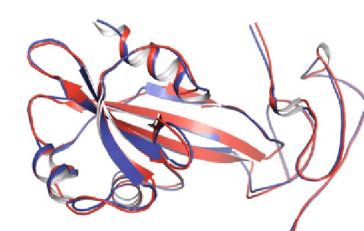
drTRPM2 1161 HRIHDTAEKVGAMSELLEREQEMVSATMAK-----RLARLEEQVSES AKALRWI IDALKSQGCKSKVQPPLMRSSSDRD-DGDSSQET- 1244
hsTRPM2 1146 QKIEDISNKVDAMVDLLDLPLKRSGSMEQ-----RLASLEEQVAQT AQALHWIVRTL RASGFSSEADVPTLASQKAAEEDAEPPGGRKKT 1231
nvTRPM2 1186 ERVRALGDRVDCINSQLNRLVDSMSGTRAHALTDGNGLEGGHDSGRLARMEVELSSNSLQKILALLQ-----QPPPVKQAAV 1266
hsNUDT9 1 MAGRLLGKALAAVSLSLALASVTIRSSRC-----RG IQA-----FRNSFSS--SW----FHLN-----TNVMSG--S 54

drTRPM2 1245 -DDEEAPMFARQLQYPDSTVRRFPVPEEKVSWEVNFSPYQPPVYNQDSSSEDTS-----ALDKHRNPGG 1309
hsTRPM2 1232 EEPGDSYVNARHLLYPNCPVTRFPVPNEKVPWETEFLLYDPFFYTAERKDAAAMD--PMGDTLEPLSTIQYN--VVDGLRDRRSFHGPYTVQAGLPLNPMG 1329
nvTRPM2 1267 PIQLTLLHYKARSSPYPGSTAKRFQVQDNMVDWQVPFPDYKPVNYTAPVVLANPVWADKDLMA MSPPELPYNQMDHTCNVNRVSYNGTYVVKDGLPLNPMG 1368
hsNUDT9 55 NGSKENSHNKARTSPYPGSKVERSQVPNEKVGWLVEWQDYKVEYTA VSVLAGPRWADPQISESNFSPKFN---EKDGHVERKSKNGLYE IENGRPRNPG 152

drTRPM2 1310 RTGIRGKGAALNTLGNHILHPIFTRWRDAEH-----KVLEFLAWWEDA EKRWALLGGPAQPDDEPLAQVLERILGKKLNEK-----TK 1386
hsTRPM2 1330 RTGLRGRGSLSCFGPNHTLYPMVTRWRNEDGAIC--RKS IKKMLEVLVVKLPLSEHWALPGGSREPGEMLPKRLKRILRQEHWS-----FE 1415
nvTRPM2 1369 RTGMQGRGLLGRFGPNHAADPVVTRWKRTSAGVML--QGGKKVLEFVAIQRKDNQWAI PGGMVPEGLVTQALKA EFGEEAMAKLNVSQEEKERIAKQIE 1467
hsNUDT9 153 RTGLVGRGLLGRWGPNHAAPIITRWKRDS SGNKIMHPVSGKHILQFVAIKRKDCGEWAI PGGMVDGGEKISATLKREFGEEALNSLQKTS AEKREIEEKLH 254

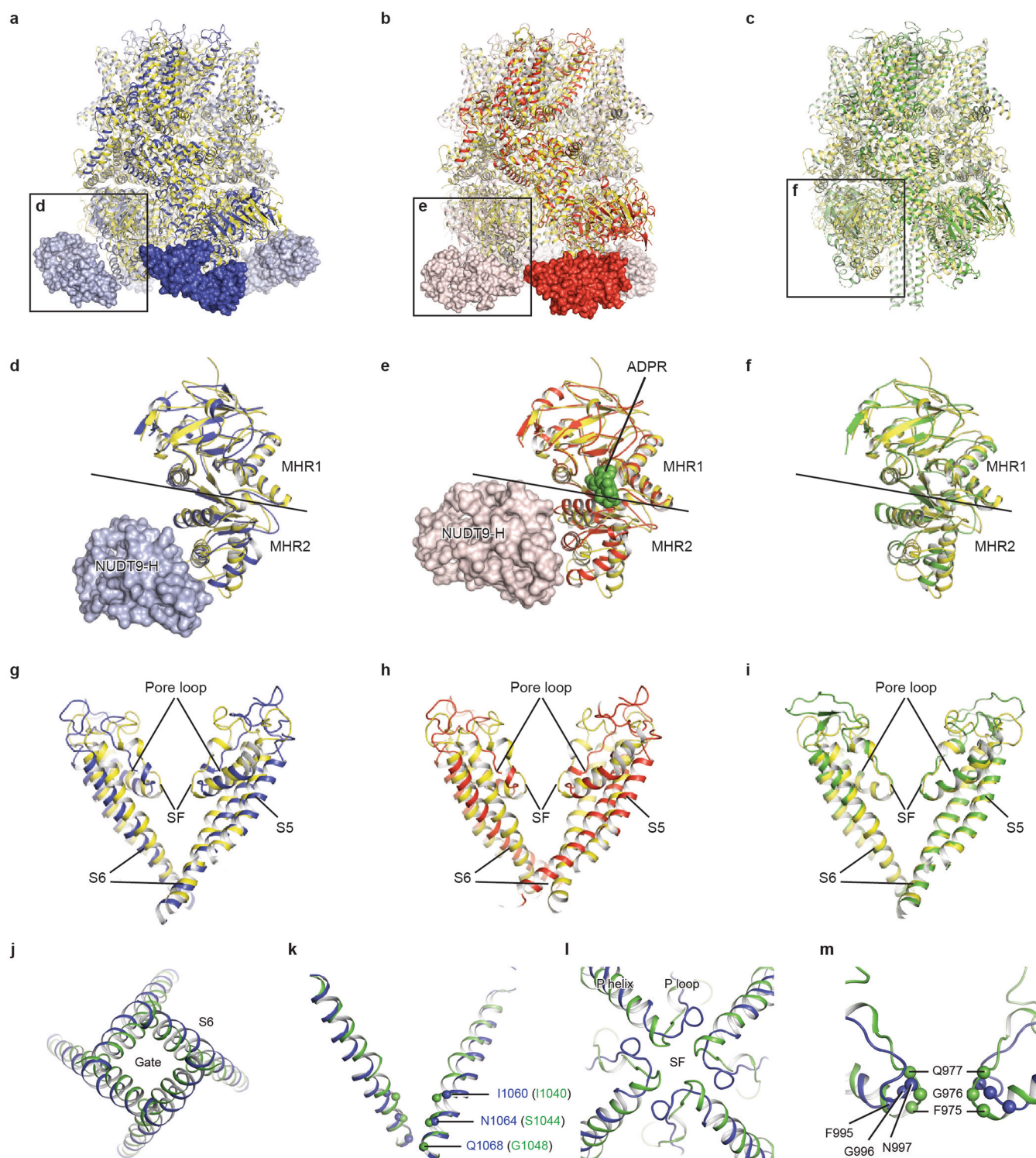
drTRPM2 1387 TLLKAG-EEVYKGYVDDSRNTDN AWEETSIITLHCDKNTPLMADLNH MVESLSHQPLQWREVSSDACRCSYQREALRQIAHHHNTYF----- 1474
hsTRPM2 1416 NLLKCG-MEVYKGYMDPRNTDN AWEETVA VSVHFQDQNDVELNRLNSNLHACDSGASIRWQVVDRIPLYANHKTLLQKAAAEFGAHY----- 1503
nvTRPM2 1468 RL FQGG-QEIKGYVDDSRNTDN AWEETVA VSNFHDDKGD L-F---GDITLQAGDDAAAVRWQRVSGNIPL YASHVSI LEKVAKMRDAAF----- 1551
hsNUDT9 255 KLFSDQHLVLYKGYVDDSRNTDN AWEETEA VNYHDETGEI-M---DNLMLEAGDDAGKVKWVDINDKLKLYASHSQFIKLVAEKRAHWS EADSEADCHAL 350

```

b**c****d****e**

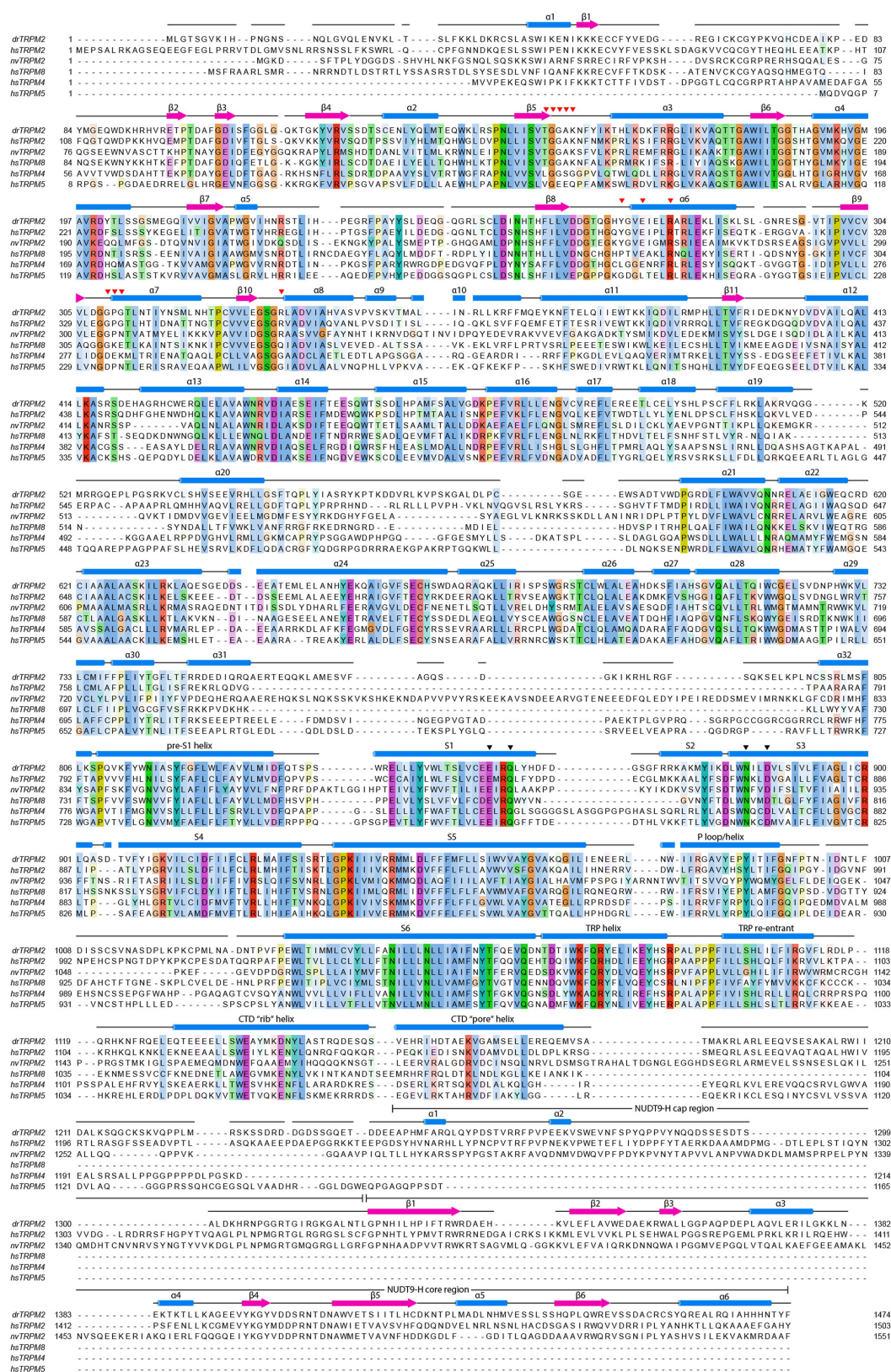
Extended Data Fig. 5 | The NUDT9-H domain. **a**, Sequence alignment of zebrafish (*Dr*), starlet sea anemone (*Nv*) and human (*Hs*) TRPM2 NUDT9-H domains with human NUDT9 using Clustal Omega⁵⁵. **b, c**, Superimposition of the zebrafish ADPR/Ca²⁺-TRPM2 NUDT9-H

domain (red) with human NUDT9 (green, PDB ID: 1Q33). Cap and core regions are indicated. **d, e**, Superimposition of the NUDT9-H domains of the zebrafish ADPR/Ca²⁺-TRPM2 structure (red) and the EDTA-TRPM2 structure (blue).



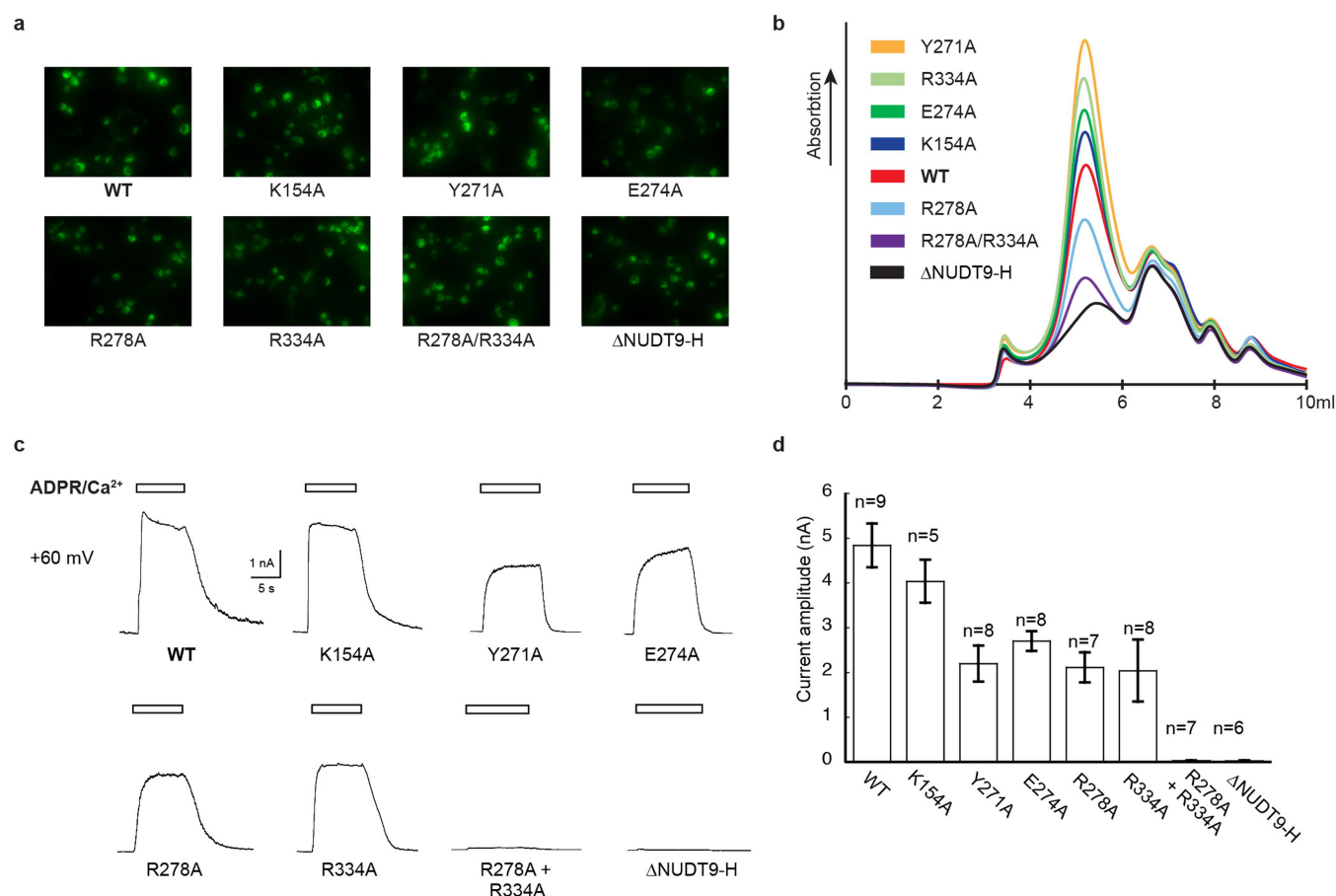
Extended Data Fig. 6 | Comparison of NvTRPM2 with DrTRPM2 (EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2) and HsTRPM4, and comparison of the gate and selectivity filter of EDTA-TRPM2 with those of HsTRPM4 (PDB: 5WP6). **a–c**, Superimposition of EDTA-TRPM2 (**a**, blue, r.m.s.d = 50 Å, overall, main chain atoms only), ADPR/Ca²⁺-TRPM2 (**b**, red, r.m.s.d = 50 Å, overall, main chain atoms only) and HsTRPM4 (**c**, green, r.m.s.d = 47.5 Å, overall, main chain atoms only) with NvTRPM2 (yellow). The NUDT9-H domain is completely invisible in NvTRPM2. **d–f**, Superimposition of the MHR1/2 domains of EDTA-TRPM2 (**d**, blue), ADPR/Ca²⁺-TRPM2 (**e**, red) and HsTRPM4 (**f**, green) with NvTRPM2 (yellow). **g–i**, Superimposition of the transmembrane domains of EDTA-TRPM2 (**g**, blue), ADPR/Ca²⁺-TRPM2 (**h**, red) and

HsTRPM4 (**i**, green) with NvTRPM2 (yellow). The transmembrane domain of NvTRPM2 is distinct from EDTA-TRPM2 and ADPR/Ca²⁺-TRPM2 but very similar to that of HsTRPM4. **j, k**, Comparison of the gates of TRPM2 (blue) and TRPM4 (green) viewed from the intracellular side of the membrane (**j**), or viewed parallel to the membrane (**k**). Only two subunits are shown in (**k**) for clarity. **l, m**, Comparison of the selectivity filters of DrTRPM2 and HsTRPM4 viewed from the extracellular side of the membrane (**l**), or viewed parallel to the membrane (**m**). Only two subunits are shown in **m** for clarity. The superimposition was performed by aligning the P loop and S6 (residues 958–1050 in HsTRPM4 and residues 978–1069 in DrTRPM2). The C α atoms of the residues in the selectivity filter and gate are shown as spheres.



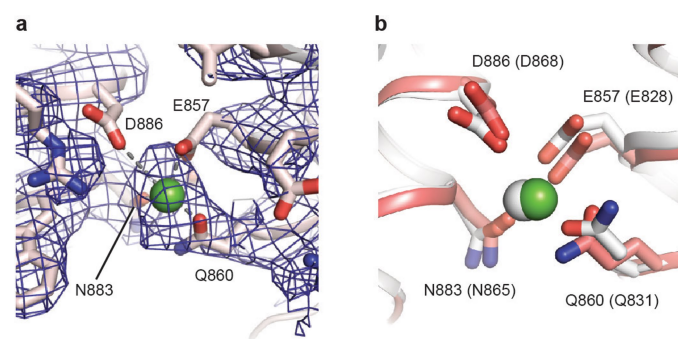
Extended Data Fig. 7 | Secondary structure arrangement of *DrTRPM2* and sequence alignment of TRPM family channels. The secondary structure prediction of *DrTRPM2* was performed using the JPred online server⁵⁶. The sequences (TRPM2 from zebrafish; TRPM2, TRPM4, TRPM5 and TRPM8 from human; TRPM2 from starlet sea anemone)

were aligned using Clustal Omega⁵⁵. Residues that are involved in ADPR binding and calcium binding are marked with red and black triangles, respectively. The cap and core regions of the NUDT9-H domain are indicated.



Extended Data Fig. 8 | Electrophysiological experiments. **a, b**, Alanine mutations were introduced within the ADPR-binding pocket and showed expression levels comparable to that of wild-type TRPM2, observed from both transfected cells (**a**) and fluorescence-detection size-exclusion chromatography (**b**). **c, d**, Electrophysiological experiments were carried out to measure the amplitude of agonist-induced current in inside-out patches pulled from HEK293 cells, with **c** showing the representative current and **d** showing the statistics of current amplitude and cell numbers. At +60 mV, robust current (4.87 ± 0.55 nA, $n = 9$ cells) could be detected when applying 0.1 mM ADPR and 1 mM Ca²⁺ onto inside-out

patches expressing wild-type TRPM2. Single mutations (K154A, Y271A, E274A, R278A and R334A) each show robust channel activation ($n = 5$ cells, 8 cells, 8 cells, 7 cells and 8 cells for the corresponding mutants) and—other than K154A receptors, which did not show a markedly reduced current amplitude—mean current amplitudes for the mutated receptors were around two- to threefold smaller compared to the wild type. Introducing double mutations R278A/R334A ($n = 7$ cells) to the receptor nearly abolished ADPR/Ca²⁺-induced current. Deletion of the NUDT9-H domain (Δ NUDT9-H) ($n = 6$ cells) also nearly abolished the ADPR/Ca²⁺-induced current. Data are shown as mean \pm s.e.m.



c

		▼	▼		▼	▼	
<i>drTRPM2</i>	857	E	I	R	Q	L	Y - - - L W N I L D 886
<i>hsTRPM2</i>	843	E	M	R	Q	L	F - - - F W N K L D 872
<i>nvTRPM2</i>	893	E	I	R	Q	L	A - - - T W N F V D 921
<i>hsTRPM8</i>	782	E	V	R	Q	W	Y - - - L W N V M D 802
<i>hsTRPM4</i>	828	E	L	R	Q	G	L - - - S W N Q C D 868
<i>hsTRPM5</i>	782	E	I	R	Q	G	F - - - N W N K C D 811
<i>hsTRPM1</i>	875	K	I	R	E	I	L - - - Y W N I T D 903
<i>hsTRPM3</i>	942	K	M	R	E	I	L - - - Y W N V T D 970
<i>hsTRPM6</i>	886	V	V	R	E	I	C - - - Y W N L T E 914
<i>hsTRPM7</i>	900	K	V	R	E	I	F - - - Y F N I S D 928

Extended Data Fig. 9 | Putative calcium-binding site. **a**, Densities of the putative Ca^{2+} -binding site and adjacent residues. **b**, Comparison of the Ca^{2+} -binding site in *DrTRPM2* (red) and *HsTRPM4* (white). Residues coordinating the Ca^{2+} ion are indicated, with residues of *HsTRPM4* shown in parentheses. **c**, Sequence alignment of the putative Ca^{2+} -binding site within the TRPM family. Residues coordinating the Ca^{2+} ion are marked with a black triangle.

Extended Data Table 1 | Statistics of 3D reconstruction and model refinement

Data collection/processing	EDTA-TRPM2	ADPR/Ca ²⁺ -TRPM2
Microscope	Titan Krios (FEI)	Titan Krios (FEI)
Voltage (kV)	300	300
Defocus range (μM)	1.0 – 2.5	1.0 – 2.5
Exposure time (s)	8	8
Dose rate (e ⁻ /Å ² /s)	6.76	6.76
Number of frames	40	40
Pixel size (Å)	1.074	1.074
Particles picked	857003	666714
Particles 2D	570680	439945
Particles refined	183041	227007
Resolution (Å)	3.8	3.3
FSC threshold	0.143	0.143
Resolution range (Å)	322.2 – 3.8	322.2 – 3.3
Model statistics		
Number of atoms	33120	35516
Protein	33120	35368
Ligand	0	148
r.m.s. deviations		
Bond length (Å)	0.01	0.008
Bond angle (°)	1.406	1.337
Ramachandran plot		
Favored (%)	94.1	95.6
Allowed (%)	5.9	4.4
Disallowed (%)	0	0
Rotamer outlier (%)	0.1	0.1
EMRinger score	2.0	2.0

RETRACTION NOTE

<https://doi.org/10.1038/s41586-018-0311-z>

Retraction Note: DDX5 and its associated lncRNA *Rmrp* modulate T_H17 cell effector functions

Wendy Huang, Benjamin Thomas, Ryan A. Flynn, Samuel J. Gavzy, Lin Wu, Sangwon V. Kim, Jason A. Hall, Emily R. Miraldi, Charles P. Ng, Frank Rigo, Sarah Meadows, Nina R. Montoya, Natalia G. Herrera, Ana I. Domingos, Fraydoon Rastinejad, Richard M. Myers, Frances V. Fuller-Pace, Richard Bonneau, Howard Y. Chang, Oreste Acuto & Dan R. Littman

Retraction to: Nature <https://doi.org/10.1038/nature16193>, published online 16 December 2015; corrected 20 January 2016.

In follow-up experiments to this Article, we have been unable to replicate key aspects of the original results. Most importantly, an RNA-dependent physical association of ROR γ t and DDX5 cannot be reproduced and is not substantiated upon further analysis of the original data. The authors therefore wish to retract the Article. We deeply regret this error and apologize to our scientific colleagues. Dan R. Littman, Benjamin Thomas, Ryan A. Flynn, Samuel J. Gavzy, Lin Wu, Sangwon V. Kim, Jason A. Hall, Emily R. Miraldi, Charles P. Ng, Frank Rigo, Sarah Meadows, Nina R. Montoya, Natalia G. Herrera, Ana I. Domingos, Fraydoon Rastinejad, Richard M. Myers, Frances V. Fuller-Pace, Richard Bonneau, Howard Y. Chang and Oreste Acuto agree to the Retraction. Wendy Huang declined to sign the Retraction letter.

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-018-0351-4>

Author Correction: Gamma frequency entrainment attenuates amyloid load and modifies microglia

Hannah F. Iaccarino, Annabelle C. Singer, Anthony J. Martorell, Andrii Rudenko, Fan Gao, Tyler Z. Gillingham, Hansruedi Mathys, Jinsoo Seo, Oleg Kritskiy, Fatema Abdurrob, Chinnakkaruppan Adaikkan, Rebecca G. Canter, Richard Rueda, Emery N. Brown, Edward S. Boyden & Li-Huei Tsai

Correction to: *Nature* <https://doi.org/10.1038/nature20587>, published online 07 December 2016.

In Extended Data Fig. 8 of this Article, we inadvertently copied the data between $A\beta_{1-40}$ levels at 1 h and $A\beta_{1-42}$ levels at 1 h. We have corrected the $A\beta_{1-42}$ graph and re-run the statistical analysis. We see a significant reduction in $A\beta_{1-40}$ and $A\beta_{1-42}$ levels 1 h after 40 Hz flicker, consistent with the other independent replications in the Article and with our immunohistochemistry analysis of $A\beta$ plaques. Our overall findings and conclusions are not changed by these results. We also accidentally omitted the raw ELISA data values for $A\beta_{1-40}$ and $A\beta_{1-42}$, for 'Dark 1hr' and '40 Hz 1 hr wait 1 hr' in Extended Data Table 1. The Supplementary Information to this Amendment contains the old, incorrect Extended Data Fig. 8 and Extended Data Table 1, for transparency. These errors have been corrected online.

Supplementary information is available for this Amendment at <https://doi.org/10.1038/s41586-018-0351-4>.

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-018-0314-9>

Publisher Correction: Deterministic delivery of remote entanglement on a quantum network

Peter C. Humphreys, Norbert Kalb, Jaco P. J. Morits,
Raymond N. Schouten, Raymond F. L. Vermeulen,
Daniel J. Twitchen, Matthew Markham & Ronald Hanson

Correction to: *Nature* <https://www.nature.com/articles/s41586-018-0200-5>, published online 13 June 2018.

In this Letter, the received date should be 20 December 2017, instead of 27 April 2018. This has been corrected online.

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-018-0355-0>

Publisher Correction: A naturally occurring antiviral ribonucleotide encoded by the human genome

Anthony S. Gizzi, Tyler L. Grove, Jamie J. Arnold, Joyce Jose, Rohit K. Jangra, Scott J. Garforth, Quan Du, Sean M. Cahill, Natalya G. Dulyaninova, James D. Love, Kartik Chandran, Anne R. Bresnick, Craig E. Cameron & Steven C. Almo

Correction to: *Nature* <https://doi.org/10.1038/s41586-018-0238-4>, published online 20 June 2018.

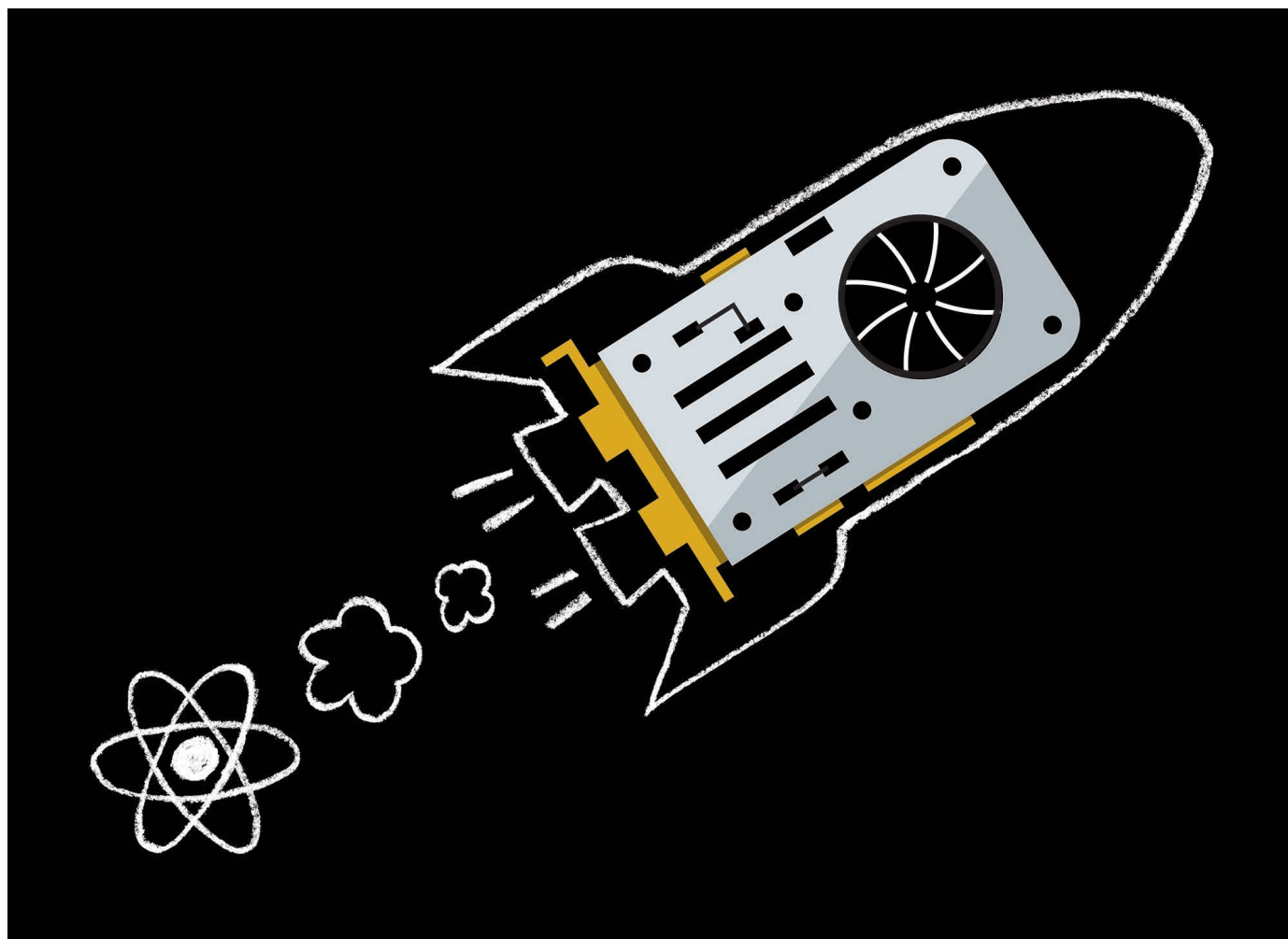
In the HTML version of this Letter, owing to a typesetter error, Extended Data Fig. 4 incorrectly corresponded to Fig. 4 (the PDF version of the figure was correct). This has been corrected online.

TOOLBOX

ACCELERATE YOUR SCIENCE WITH GRAPHICS CARDS

Graphics processing units aren't just of interest to gamers and cryptocurrency miners — parallel processing is increasingly being used to turbocharge scientific research.

ILLUSTRATION BY THE PROJECT TWINS



BY DAVID MATTHEWS

Evan Schneider, an astrophysicist at Princeton University in New Jersey, spent her doctorate learning to harness a chip that's causing a quiet revolution in scientific data processing: the graphics processing unit (GPU).

Over the past decade or so, GPUs have been challenging the conventional workhorse of computing, the central processing unit (CPU), for dominance in computationally intensive

work. As a result, chips that were designed to make video games look better are now being deployed to power everything from virtual reality to self-driving cars and cryptocurrencies.

Of the 100 most powerful supercomputing clusters in the world, 20 now incorporate GPUs from NVIDIA, one of the leading chipmakers. That includes the world's fastest computer, the Summit cluster at the US Department of Energy's Oak Ridge National Laboratory in Tennessee, which boasts more than 27,000 GPUs.

For Schneider, running astrophysical

models on GPUs "has opened up a new area of my work that just wouldn't have been accessible without doing this". By rewriting her code to run on GPU-based supercomputers rather than CPU-focused ones, she has been able to simulate regions of the Galaxy in ten times as much detail, she says. As a result of the increase in resolution, she explains, the entire model now works differently — for example, giving new insights into how gas behaves on the outskirts of galaxies.

Put simply, GPUs can perform vastly ►

► more calculations simultaneously than CPUs, although these tasks have to be comparatively basic — for example, working out which colour each pixel should display during a video game. By comparison, CPUs can bring more power to bear on each task, but have to work through them one by one. GPUs can therefore drastically speed up scientific models that can be divided up into lots of identical tasks, an approach known as parallel processing.

There are no hard rules about which types of calculation GPUs can accelerate, but proponents agree they work best when applied to problems that involve lots of things — atoms, for example — that can be modelled simultaneously. The technique has become particularly prevalent in fields such as molecular dynamics, astrophysics and machine learning.

Schneider, for example, models galaxies by splitting them up into millions of discrete areas and then dividing the work of simulating each of them between a GPU's many cores, the units that actually execute calculations. Whereas CPUs normally have at most tens of cores — the number has increased as they themselves have become more parallel — GPUs can have thousands. The difference, she explains, is that whereas each CPU core can work autonomously on a different task, GPU cores are like a workforce that has to carry out similar processes. The result can be a massive acceleration in scientific computing, making previously intractable problems solvable. But to achieve those benefits, researchers will probably need to invest in some hardware or cloud computing — not to mention re-engineering their software.

A BIT AT A TIME

Fundamentally, parallelizing work for GPUs means breaking it up into little pieces and farming those pieces out to individual cores, where they are run simultaneously. There are several ways scientists can get started. One option, Schneider says, is to attend a parallel-processing workshop, where “you’ll spend literally a day or two with people who know how to program with GPUs, implementing the simplest solutions”.

OpenACC, for instance, is a programming model that allows scientists to take code written for CPUs and parallelize some processes; this allows researchers to get a feel for whether their code will run significantly faster on GPUs, Schneider says. The OpenACC community runs regular workshops and group programming events called hackathons to get scientists started on porting their code to GPUs. When contemplating whether to experiment with GPUs, “the first thing to think about is: ‘Is there a piece of my problem where everything could be done in parallel?’”, says Schneider.

The next step is to rewrite code specifically to take advantage of GPUs. To speed up his code, Philipp Germann, a systems biologist at the European Molecular Biology Laboratory in

Barcelona, Spain, learnt CUDA, an NVIDIA-created parallel processing architecture that includes a language similar to C++ that is specifically designed for GPUs. “I’m definitely not a particularly experienced programmer,” he admits, but says the tool took only around two weeks to learn.

One cell-modelling program that Germann and his colleagues created proved to be two to three orders of magnitude faster on GPUs than were equivalent CPU-based programs. “We can really simulate each cell — when will it divide, how will it move, how will it signal to others,” Germann says. “Before, that was simply not possible.”

CUDA works exclusively on NVIDIA chips; an alternative tool, the open-source OpenCL, works on any GPU, including those from rival chipmaker AMD. Matthew Liska, an astrophysicist at the University of Amsterdam, prefers CUDA for its user-friendliness. Liska wrote GPU-accelerated code simulating black holes as part of a research project (M. Liska *et al. Mon. Not. R. Astron. Soc. Lett.* **474**, L81–L85; 2018); the GPUs accelerated that code by at least an order of magnitude, he says. Other scientists who spoke to *Nature* also said CUDA was easier to use, with plenty of code libraries and support available.

You might need to block off a month or two to focus on the conversion, advises Alexander Tchekhovskoy, an astrophysicist at Northwestern University in Evanston, Illinois, who was also involved in Liska’s project. Relatively simple code can work with little modification on GPUs, says Marco Nobile, a high-performance-computing specialist at the University of Milan Bicocca in Italy who has co-authored an overview of GPUs in bioinformatics, computational biology and systems biology (M. S. Nobile *et al. Brief. Bioinform.* **18**, 870–885; 2017). But, he warns, for an extreme performance boost, users might need to rewrite their algorithms, optimize data structures and remove conditional branches — places where the code can follow multiple possible paths, complicating parallelism. “Sometimes, you need months to really squeeze out performance.”

And sometimes, the effort simply isn’t worth it. Dagmar Iber, a computational biologist at the Swiss Federal Institute of Technology in Zurich, says that her group considered using GPUs to process light-sheet microscopy data. In the end, the researchers managed to get acceptable results using CPUs, and decided not to explore a GPU acceleration because it would have meant making too many adaptations. And not all approaches will work, either: one of Tchekhovskoy’s attempts yielded no speed improvement whatsoever over a CPU. “You can invest a lot of time, and you might not get much return on your investment,” he says.

More often than not, however, that’s not the case; he says that in the field of fluid dynamics, researchers typically have seen speed-ups “from a factor of two to an order of magnitude”.

CLOUD PROCESSING

As for the hardware itself, all computers need a CPU, which acts as the computer’s brain, but not all come with a dedicated GPU. Some instead integrate their graphics processing with the CPU or motherboard, although a separate GPU can normally be added. To add multiple GPUs, users might need a new motherboard with extra slots, says Liska, as well as a more robust power supply.

One way to test out parallel processing without having to actually buy new hardware is to rent GPU power from a cloud-computing provider such as Amazon Web Services, says Tim Lanfear, director of solution architecture and engineering for NVIDIA in Europe, the Middle East and Africa. (See, for example, this computational notebook, which exploits GPUs in the Google cloud: go.nature.com/2ngfst8.) But cloud computing can be expensive; if a researcher finds they need to use a GPU constantly, he says, “you’re better off buying your own than renting one from Amazon”.

Lanfear suggests experimenting with parallel processing on a cheaper GPU aimed at gamers and then deploying code on a more professional chip. A top-of-the-range gaming GPU can cost US\$1,200, whereas NVIDIA’s Tesla GPUs, designed for high-performance computing, have prices in the multiple thousands of dollars. Apple computers do not officially support current NVIDIA GPUs, only those from AMD.

Despite the growth of GPU computing, the technology’s progress through different scientific fields has been patchy. It has reached maturity in molecular dynamics, according to Lanfear, and taken off in machine learning. That’s because the algorithm implemented by a neural network “can be expressed as solving a large set of equations”, which suits parallel processing. “Whenever you’ve got lots of something, a GPU is typically good. So, lots of equations, lots of data, lots of atoms in your molecules,” he says. The technology has also been used to interpret seismic data, because GPUs can model millions of sections of Earth independently to see how they interact with their neighbours.

But as a practical matter, harnessing that power can be tricky. In astrophysics, Schneider estimates that perhaps 1 in 20 colleagues she meets have become adopters, held back by the effort it takes to rewrite code.

Germann echoes that sentiment. “More and more people are recognizing the potential,” he says. But, “I think still quite a few labs are scared to some degree” because “it has the reputation of being hard to program”. ■

David Matthews is a freelance writer based in Berlin.

CAREERS

CANADA Visa regulations might be pushing out postdoctoral researchers **p.155**

SUPPORT Female peers boost likelihood of science-PhD completion for women **p.155**

FACEBOOK Follow us for career listings and advice www.facebook.com/naturejobs

JANNES DE VILLIERS



Shivan Parusnath at the University of the Witwatersrand in Johannesburg, South Africa, thinks that media interviews can boost employment chances.

MEDIA

Smile for the camera

Use interviews to promote your science, raise your profile and practise your media skills.

BY AMBER DANCE

The phone rings in Chwee Teck Lim's office at the National University of Singapore. He answers enthusiastically, yet with a bit of apprehension. On the other end of the line is a reporter, eager to hear about the biomedical engineer's latest research. But will she describe his work accurately? Will she give credit to his collaborators? Lim knows from experience that the final article or broadcast probably won't come out the way he expects.

But so long as the essential points of Lim's work are covered correctly, he's happy. "They are doing me a favour by publicizing my research," he points out.

Lim, acting director of the Biomedical Institute for Global Health Research and

Technology at his university, also considers it a responsibility to answer those calls. They provide opportunities to inform taxpayers about publicly funded research. In today's anti-science climate, with some politicians denying climate change and some parents eschewing life-saving vaccines, that's particularly important, says David Shukman, science editor at BBC News in London. "We're at a time when truth is at a premium, where facts need to be established and form the basis of public policy debate," he says. "Science plays a crucial role in that, and I think the key mechanism is scientists explaining those truths, those facts, to a wider audience."

Preparation is crucial: to effectively convey your message, you need to define it first. It's also important to understand the media outlet and the needs of its audience, and to recognise that

you will surrender control of the final product to the reporter or producer. For those who are new to being interviewed or being on camera, training courses — and simple practice — can make the experience easier, even enjoyable.

Giving interviews can benefit a research programme, too. The press office at Imperial College London has informally tracked the outcomes of Imperial scientists' media contact. Interviewees report more citations for their work, contacts from potential collaborators and invitations to speak at conferences, says Laura Gallagher, head of news and media at Imperial. Scientists who speak to the media might be approached by industry investors, philanthropic donors or volunteers who are eager to participate in clinical trials.

Dean Falk, an evolutionary anthropologist ►

► at Florida State University in Tallahassee, has a reporter to thank for a valuable collaboration. In 1994, Associated Press science writer Malcolm Ritter contacted her about a paper written by researchers in Austria on radio imaging of Ötzi, the iceman found mummified in the Austrian–Italian Alps (D. zur Nedden *et al. Radiology* **193**, 269–272; 1994). Falk praised the paper, then forgot about it.

Later, Ritter wrote back. One of the study's authors, Horst Seidler at the University of Vienna, wanted Falk's contact details so that he could invite her there. That led to a long-standing collaboration: the researchers co-wrote papers; Falk joined the team on field expeditions in Ethiopia; and she gained an honorary appointment at the University of Vienna.

Shivan Parusnath, a graduate student in zoology at the University of the Witwatersrand in Johannesburg, South Africa, thinks that media interviews boost his employment chances by publicizing his name and accomplishments. And talking to the media has helped his research, too. He studies a vulnerable lizard called the sungazer (*Smaug giganteus*), and media coverage of his work has led farmers to contact him about sungazers living on their property. They invite him to visit and take samples for his DNA database.

Parusnath delights in thinking about future scientists reading or hearing about his research. He recalls listening to the radio from the back seat of his parents' car as a child. When he gave a live radio interview, Parusnath says, "I was just thinking about a little 'me' listening somewhere at home, maybe getting excited about it."

CONVERSATION PREPARATION

Before any interview, Parusnath and other media-savvy scientists prepare. They find out what the article or programme will be about, and distil what they want to say into two or three key messages. "When the interviewer asks you a question, mentally run through your list to see if any of those points can serve as an answer," says Sabrina Stierwalt, an astrophysicist at NASA and the California Institute of Technology in Pasadena. "You'll be less likely to get sidetracked." Scientists can come up with analogies and examples to use instead of technical jargon, and it's also helpful to understand basic media concepts such as 'off the record' (see 'What interviewees need to know').

Reporters might not provide their specific questions ahead of time, but they can usually offer some idea of their topic or angle, says Sanam Mustafa, a molecular pharmacologist at the University of Adelaide in Australia. If they won't, something's fishy.

For example, she underwent media training as part of her participation in the 'Superstars of STEM' programme, which publicizes women in science and technology in Australia. Just before International Women's Day on 8 March this year, a controversial television show contacted the programme's media officer, looking for interviewees. The producers promised

a 'positive' story, but when the media officer pressed for more details, they wouldn't say anything else, and rescinded the invitation.

Assuming that a researcher is ready to trust an interviewer, what can one expect? Science reporters often ask the same kinds of question. The non-profit organization Sense About Science USA surveyed 218 science journalists in 2015 and listed questions that come up often, including: 'How was a study conceived or structured?', 'How were the conclusions reached?', 'What do the findings mean in the context of the field?' and 'What unknowns remain?'

Natalie Hodgson, a media manager at the Wellcome Trust, a biomedical research charity in London, adds one more key question to consider: "What is the headline that you wouldn't want to see?" Thinking about that hypothetical horror helps scientists to focus on clear explanations and remember to bring up any caveats about their work, she says.

Tara Shears, a particle physicist at the University of Liverpool, UK, likes to spend 20 or 30 minutes before an interview writing down what she plans to say about her work at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. For her, research on antimatter is about understanding the nature of the Universe. But she realizes that might be rather abstract for a commuter skimming the news on the train.

To relate antimatter to everyday life, Shears often turns to what she calls her two "golden fallbacks". One is to point out that medical PET (positron emission tomography) scanners work only because of antimatter: the radioactive tracer emits the antimatter version of an electron, and its destructive clashing with a

regular electron creates the signal that the scanner reads.

Shears also likes to mention that bananas emit antimatter, because the fruit contains a radioactive isotope of potassium. "It's not dangerous," she hastens to add.

Lim makes sure to prepare resources that might help the reporter. These could include a live demonstration, slides showing samples or prototypes, images with copyright information and the names of other scientists who could objectively comment on his work.

A MATTER OF TRUST

But no preparation will give researchers control over the final piece. Scientists can ask to see the final copy before it airs or goes to press, but the answer will probably be no. According to the Sense About Science survey, some reporters will occasionally send the relevant portion of an article or a scientist's quotes. But most science journalists never send the entire piece for an interviewee to check.

Why not? One reason is practical. Journalists often work to tight deadlines, putting the finishing touches to stories right before they go live. There isn't time to track down all the scientists again.

The other reason is ethical. "It's a basic journalistic tenet that subjects don't have editorial control over the product," says Bruce Mohun, a television science journalist in Vancouver, Canada. Most political journalists would never show the president or prime minister their work before publishing it; science journalists work in the same way.

Reporters will return to interviewees to clear up a point or check a fact, but errors can and do

KEY TERMS

What interviewees need to know

Here are important terms and tenets to remember about interviews with journalists.

● **On the record.** Anything you say to a reporter is, by default, on the record and can be attributed to you.

● **Off the record.** Nothing from this conversation can be published. The journalist must agree to these terms before the discussion. Some media professionals caution that it's safest to assume that everything is on the record, however, and to speak accordingly.

● **On background.** This also requires an agreement between the reporter and source. It may mean that the information is publishable, but cannot be attributed to you, or that you can only be described in vague terms, such as 'a government researcher'.

● **Embargo.** When a paper is coming out in a scientific journal, those findings are considered embargoed — temporarily restricted from being published elsewhere.

Reporters don't release news of the study until a date set by the journal. In return, they get advance access. It's fine to speak with journalists before the embargo date, but it never hurts to remind them of the embargo.

Results presented at large scientific meetings are also fair game for news coverage, and reporters are usually aware that data are preliminary or not yet peer-reviewed. Journals vary in their proscriptions for presenting scientists. For example, *Science* and the *Journal of the American Medical Association* say that scientists can talk to journalists, but should limit their conversation to what was in the presentation.

As for talking to reporters about unrepresented results well before publication or even submission, again, policies vary. When in doubt, consult with press officers from your institution or the journal for guidance. **A.D.**



Sanam Mustafa at the University of Adelaide in Australia did media training to prepare for interviews.

creep in. Journalists will be eager to correct factual mistakes, says Ritter. “I think we would also take a serious look if there was not a factual error, but if we gave a wrong impression about something,” he says.

Other changes to published work are less likely to happen. “Where reporters are going to be less sympathetic is if you just want to change your quote, or if a reporter has skipped some detail,” says Valerie Jamieson, creative director of the UK *New Scientist Live* exhibition. Writers and producers might leave out information that scientists think is important.

For example, Parusnath once spent four days in the field with a member of 50/50, a long-running environmental TV programme in South Africa. In between catching sungazers and talking about how people affect the lizards, Parusnath was careful to mention his funding sources and collaborators.

But those names didn’t make it into the documentary. One of Parusnath’s supervisors was irate, thinking he hadn’t bothered to bring up his university funding. “I have no control over whether that gets in,” Parusnath pointed out. “It probably wasn’t relevant to the story for the people producing it.”

Jamieson explains that long lists of collaborators or funding sources, or detailed job titles, simply aren’t interesting for readers or viewers. And with only a few hundred words or a couple of minutes in which to share the key points of a study, there just might not be room for those details.

Lim has a strategy to share the limelight with his collaborators: he invites them to the interview. He’s an engineer, but if the study has clinical implications, he’ll ask a clinician to join in to answer questions Lim can’t.

Sometimes scientists might be surprised to find that they don’t make an appearance in the final article or broadcast at all. That doesn’t mean the interview was pointless,

says Ritter. “Even if a scientist is not quoted in a particular story, whatever they tell us is helping us shape that story.”

If that all sounds daunting, there are ways to get better at interviewing. Press officers can run a mock interview to help scientists warm up. And training courses can also help researchers to gain confidence, says Shears. She took a one-day workshop offered by the Royal Society in London to practise on-camera work. “It was the most excruciating day of my whole life,” recalls Shears, who did three practice interviews and then watched them with the other students. But she learnt a lot — including her tendency to avoid focusing on the camera when asked to, which she has since corrected.

Being on the radio or TV certainly adds an extra layer of complexity to interviews, although the basic tenets of preparation are the same as for written articles. It’s particularly important for novices to practise what they want to say if they’ll be on air live, with only one chance to get the story right.

For radio interviews, a researcher might have to go into the studio, or the producer might be able to record the scientist by phone if a high-quality landline is available. For television, researchers should plan ways to show the camera what they’re doing. For example, if a mathematical formula is key to the research, the producer could film the researcher writing it on a whiteboard, suggests Mohun.

Mustafa has learnt not to worry as much as she used to. “I think sometimes we can be our own worst critic,” she says, recalling her first radio interview, when she thought she bombed. But when she listened to the programme, it was good. “Every time you do an interview, you will get better at it,” she says. ■

Amber Dance is a freelance journalist in Los Angeles, California.

CANADA

Postdoc visa woes

International postdocs in Canada say that the country’s visa and immigration requirements are making it hard for them to complete their programmes and could bar them from becoming permanent residents, a report from the Canadian Association of Postdoctoral Scholars (CAPS) finds (see go.nature.com/visas). The study, based on a 2016 survey of 2,109 current and former postdocs from across Canada, documented immigration-related complaints from international researchers. More than 40% of postdocs from other countries listed “visa/work permit issues” as a major challenge. Respondents said they must reapply annually for complicated work permits, and that their institutions offered little help when confusion or questions arose. The system needlessly complicates the lives of international postdocs, and could keep some from staying in the country, says the report’s author, Joe Sparling, who chairs CAPS. In the survey, 29% of respondents said they were in the country on work permits, down from 38% in 2013. Sparling ties the decline in part to Canada’s Express Entry immigration programme. The scheme, which began in January 2015, makes it difficult for postdocs to document enough work experience to apply for permanent residency.

PHD PROGRAMMES

Support in numbers

Women who enter a US PhD programme in a science, technology, engineering or mathematics (STEM) field are less likely to graduate if relatively few other women also join, according to a report by economists Valerie Bostwick and Bruce Weinberg of Ohio State University in Columbus (see go.nature.com/2mubhhs). The authors looked at data for 2,541 students starting PhDs at public universities in Ohio from 2005 to 2009. Women accounted for nearly 40% of the sample, but their numbers varied widely between programmes. When a cohort contained just one woman, she was 12% less likely to graduate within 6 years than were her male peers. But as the proportion of women increased, so did each woman’s likelihood of obtaining a degree. The authors suggest that women’s chances of earning a STEM PhD are linked to the ‘female-friendliness’ of that programme. “If there are few or no other women in your incoming class, it can make it more difficult to complete your degree,” says Bostwick.

FOCAL POINT ON KOBE

PRODUCED IN PARTNERSHIP WITH THE FOUNDATION FOR BIOMEDICAL RESEARCH AND INNOVATION

CO-LOCATED FOR COLLABORATION

The results from **KOBE'S BIOMEDICAL HUB** show there is strength in sharing research infrastructure.

A condition known as vocal fold scarring stiffens the vocal cords and makes it difficult for people to speak or sing. It generally happens to people who use their voice a lot, such as singers, but who and why it will strike is difficult to predict. A research institute in Kobe, Japan, helped advance a first-of-its-kind clinical trial that showed how injections of growth factor can regenerate the mucosa lining of the vocal cords. The findings could help patients regain their voices.

The institute behind the study, the Translational Research Center for Medical Innovation (TRI), is one of about 350 research centres, specialized hospitals, companies and universities grouped in a unique science campus on Port Island off Kobe. Founded after the 1995 earthquake that devastated the region, the Kobe Biomedical Innovation Cluster (KBIC) is now producing some of the most cutting-edge medical research in the world.

Offshore innovation

KBIC's mission was not only to create jobs and revitalize the local economy, but to promote the health and welfare of local people and to improve medical standards across Asia. It contains facilities across the entire range of medical research and development, from basic research to clinical applications and mass production. The core resources are used by groups in and outside Kobe, including TRI, the International Medical Device Alliance, the Kobe Hybrid Business Center, and the Kobe Medical Device Development Center.

AN ISLAND OF OPPORTUNITY

Kobe Port Island in Kobe Harbour, is a man-made structure completed in 1981. An addition to the Kobe Biomedical Innovation Cluster, it houses six universities, shipping and cruise-liner docking facilities, and a zoo. It is also the stepping stone between Kobe Airport and the city of Kobe, accessed by monorail.



THE 1995 GREAT HANSHIN EARTHQUAKE
measured 6.9 on the moment magnitude scale and lasted 20 seconds.



THE KOBE BIOMEDICAL INNOVATION CLUSTER
contributed an estimated 153.2 billion yen (US\$1.37 billion) to the Japanese economy in 2015.



(Top left) © koksikoks, gyro/iStock /Getty Images Plus, (Bottom left) © cogal/Getty Images, (Top right) © Gannet77/Getty Images, (Bottom right) © Boobigum/iStock /Getty Images Plus

© gyro/iStock /Getty Images Plus

The central pillar of the cluster is the Foundation for Biomedical Research and Innovation at Kobe (FBRI). Founded in March 2000, its mission is to promote advanced clinical research, next-generation healthcare systems and collaborations among KBIC entities.

"The FBRI has been endeavouring to build an ecosystem of healthcare innovation, through engaging in R&D, supporting clinical research, providing business support and strengthening the research and business networks within the KBIC," says FBRI President Tasuku Honjo, a Kyoto University immunologist who is renowned for identifying a protein known as Programmed Cell Death Protein 1.

New medical frontiers

The FBRI is supporting a promising field of inquiry that has made headlines around the world in recent years. In 2006, Kyoto University researcher Shinya Yamanaka showed how adult cells could be converted to stem cells with the ability to differentiate into any kind of cell. These induced pluripotent stem (iPS) cells allowed researchers to avoid the ethical problems associated with using embryonic stem cells and opened up the new field of regenerative medicine. Over the past few years, researchers at the RIKEN Center for Developmental Biology (CDB) in Kobe and Kobe City Medical Center General Hospital have brought the technology to clinical application. They have been transplanting retinal cells derived from iPS cells into patients with age-related macular degeneration, a leading cause of vision loss in people over 50.

"We are now completing follow-up checks of the patients and we look forward to a new trial involving photoreceptor cells to treat retinitis pigmentosa," says Masayo Takahashi, a project leader at RIKEN CDB. "I'm sure regenerative medicine will be a very standard form of treatment in the future."

Takahashi says her work has benefited from the FBRI because it has facilitated many industry collaborations related to gene therapy.

One example is her appointment as special advisor to Healios KK, a biotech startup based in Kobe and Tokyo that is working on iPS cell clinical applications in conjunction with Sumitomo Dainippon Pharma and RIKEN.

Building next-generation tools

RIKEN hosts one of the world's most advanced computer research facilities, located on Port Island. The RIKEN Center for Computational Science (CCS) is home to the K computer, a supercomputer built by Fujitsu that was ranked fastest in the world by the TOP500 computing project in 2011; it was also the world's first supercomputer to achieve 10.51 quadrillion floating point operations per second.

Since then, the K computer has been used for everything from industrial design simulations to global climate modelling and analysis of genetic and drug data. Scientists at RIKEN and its partners are now building a post-K computer, which will be available for public use around 2021.

"The development is going extremely well, and we believe the Post-K inherits all the good traits of K and fixes its shortcomings, making it a landmark, game changing machine of the time," says Satoshi Matsuoka, a professor of computer science at Tokyo Institute of Technology and director of RIKEN CCS. "We already have the first chip and it is performing as expected, in some cases better. The chip will outclass every existing CPU by several factors in all metrics."

The KBIC and FBRI will continue to build on their legacy of making Kobe a leading biomedical cluster as they tackle the social problems of the twenty-first century.

"The FBRI will continue to carry out fundamental research mainly in fields where new medical treatments or medications are urgently required, such as cancer immunology, ageing and dementia," says Honjo. "We're focused on proposing solutions to problems along the way to realizing a society where people can enjoy good health and longevity." ■

MACULAR DEGENERATION
is a form of blindness caused by dysfunction in the retinal pigment epithelium



"I'm sure regenerative medicine will be a very standard form of treatment in the future."

The seismic rise of Kobe's biomedical hub

An ambitious plan conceived in the wake of the Great Hanshin earthquake led to **THE DEVELOPMENT OF A WORLD-CLASS BIOMEDICAL HUB.**

In 1995, a massive earthquake levelled the port city of Kobe. Barely three years later, municipal officials hatched a plan to reinvent the local economy. Their idea: transform an artificial island in Kobe harbour from a disaster zone into one of the world's foremost epicentres for biomedical research and pharmaceutical innovation.

It was an audacious proposal. The city, long reliant on heavy industry, had almost no background in the healthcare sector. And the chosen spot, Port Island, contained little more than temporary housing. But thanks to visionary leaders and strong government support, the ambitious overhaul has been an unequivocal success. "It has gone well beyond the initial plan developed 20 years ago," notes Kobe mayor, Kizō Hisamoto.

A unique hub for the life sciences

Port Island now hosts one of the largest concentrations of medical research facilities in the world. Named the Kobe Biomedical Innovation Cluster (KBIC), the life sciences hub contains around 350 companies and non-profit organizations,

including those operating several hospitals, colleges and research institutes. The hub now provides work for around 10,000 highly skilled people in biomedical treatment, research and technology. Members of KBIC plan to celebrate these achievements at a 20th anniversary commemorative ceremony and symposium on 19 October 2018.

NO TWO RESEARCHERS ARE MORE THAN A HALF-HOUR STROLL APART

The Kobe cluster stands apart from other centres of biomedical research and development. Unlike the life science hubs of North America, Europe or elsewhere in the Asia Pacific region, KBIC did not grow up around existing academic institutions and infrastructure. "Kobe began its project from almost nothing after the earthquake," Hisamoto explains. That blank slate gave KBIC organizers the freedom to design the area for optimal scientific creativity and productivity.



Aerial view of Kobe Biomedical Innovation Cluster.
The centre of Kobe city is over the red bridge at the top left of the photograph.

The heart and centre of the campus is a biotechnology area designed to unite fundamental research with clinical applications of regenerative treatments and medical devices. On one side of this scientific core lie state-of-the-art medical facilities that can accommodate around 1,500 patients, allowing clinicians to run large clinical trials while promoting new systems for routine care. On the other side lie buildings devoted to computational modelling. One of them is home to the K computer, the world's highest performance supercomputer, which accelerates drug discovery by running complex simulations.

Notably, all these resources and facilities exist within a

compact area in which no two researchers are more than a half-hour stroll apart. "That considerably facilitates information exchange and networking among the Kobe cluster," Hisamoto says. "It's a unique research and operating environment." Ease of interaction between researchers has rapidly yielded groundbreaking scientific discoveries, regenerative therapies and medical devices.

World-leading medical treatments

The most prominent achievement came in 2014 when an elderly woman suffering from age-related macular degeneration became the first person in the world to

receive tissue derived from her own induced pluripotent stem (iPS) cells. She was treated at the Kobe City Medical Center General Hospital with a therapy developed by RIKEN scientists at a nearby institute now known as the Center for Biosystems Dynamics Research.

In 2017, the same KBIC-based research team completed another world first when they implanted iPS-derived cells from an anonymous donor into a man with the same vision-destroying eye disease. Researchers expect to see other stem-cell-based therapies used in clinical practice soon for knee cartilage injuries, vocal cord scarring and blood vessel obstructions in the legs, among other disorders.

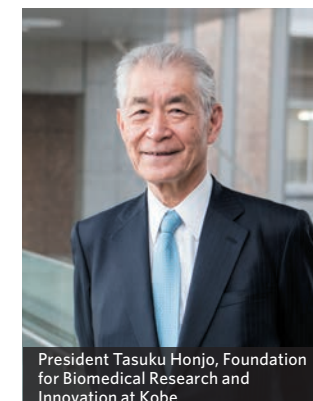
Two critical location decisions

Key to these recent accomplishments were two events in the cluster's early development. Most crucial was a decision in 2000 by RIKEN, Japan's leading basic research institute, to establish a centre for developmental and regenerative biology in Kobe. Other RIKEN units soon followed, one dedicated to molecular imaging studies and another to computational science.

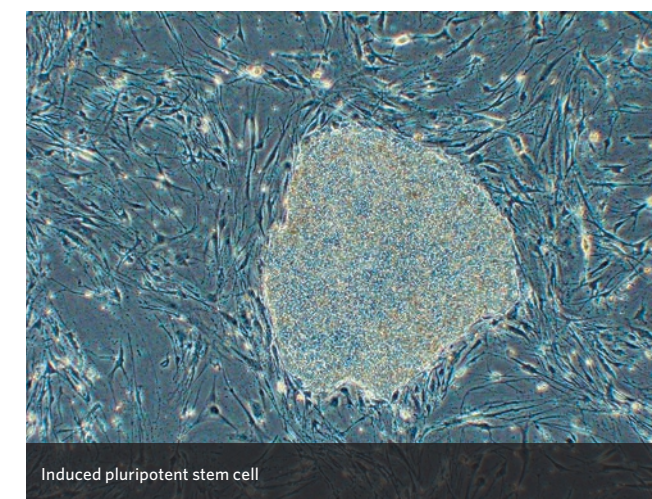
The Foundation for Biomedical Research and Innovation (FBRI) was created to facilitate clinical research and serve as a liaison between laboratory scientists, doctors and companies. The foundation opened a branch right next to the



Kobe Mayor Kizō Hisamoto



President Tasuku Honjo, Foundation for Biomedical Research and Innovation at Kobe



Induced pluripotent stem cell

©RIKEN

RIKEN research labs, allowing preclinical investigations and human trials to take place in quick succession.

In April 2018, the foundation relaunched itself with an increased focus on providing business support to strengthen public-private-academia partnerships and to promote international alliances. Part of the goal, Hisamoto explains, is for KBIC to "strengthen its efforts towards supporting life science start-ups". The cluster has done a good job of attracting existing companies, both large and small, to Kobe. FBRI leaders also hope to foster a greater spirit of high-risk, high-reward entrepreneurialism among young researchers and

encourage them to break out of the lab to form new companies.

Recalling what the city looked like 20 years ago, "it would have been hard to imagine the present-day KBIC," Hisamoto says. Given the pace of scientific achievement, the critical mass of companies and the continued growth of the cluster, it may be equally hard to picture KBIC two decades in the future. ■



Foundation for Biomedical Research and Innovation at Kobe (FBRI)
Tel: +81-78-306-2231
Email: kbic-shisatsu@fbri-kobe.org
Web: www.fbri-kobe.org/english

The next generation of superfast supercomputers

The Post K computer, a next-generation supercomputer tasked with solving some of the world's most pressing problems, is being developed by **JAPAN'S RIKEN CENTER FOR COMPUTATIONAL SCIENCE**

In June 2011, Japan's K computer became the world's fastest supercomputer, achieving a speed of over 8 petaflops — or 8 quadrillion (8 million billion) operations a second. In November the same year, the K computer — a play on the Japanese word *kei* (京), which means 10 quadrillion — attained speeds exceeding 10 petaflops, and in 2012 started full-scale operation. It has since been used to solve problems from estimating damage by earthquakes and tsunamis to discovering new drugs.

Led by the RIKEN Center for Computational

Science (R-CCS) in partnership with Fujitsu, the Flagship 2020 Project was launched by the Japanese government in 2014. It was tasked with developing the Post K computer, a next-generation supercomputer that will boast as much as 100 times the performance of its predecessor.

"With the Post K computer, we're aiming to boost computing capabilities by orders of magnitude," explains Satoshi Matsuoka, the director of the R-CCS and Japan's foremost authority on supercomputers. "We're also striving to provide enhanced

ease of use and synergy with other information technology ecosystems, such as big data and artificial intelligence."

Japan's flagship centre for high-performance computing and the developer of the K computer, R-CCS is one of the world's leading research centres for supercomputers, which have become essential tools in modern scientific research.

"WITH THE POST K COMPUTER, WE'RE AIMING TO BOOST COMPUTING CAPABILITIES BY ORDERS OF MAGNITUDE"

Matsuoka sees the future direction for high-performance computing research at R-CCS as "the science of computing — computing itself as a scientific target of investigation; science by computing, focused on applications that use the high-performance capabilities of supercomputers; and science for computing, to explore how other scientific disciplines can contribute to the development of advanced supercomputers."

Scheduled to begin full operation in 2021, the

Post K computer will tackle priority areas across a range of social and scientific fields.

Comprising hundreds of thousands of central processing units — electronic circuits that perform the instructions encoded in computer programs — that work in parallel to execute hundreds of quadrillion calculations a second, the Post K computer will also be more energy efficient, with advanced software that will allow a wide range of users to access its capabilities.

"This will allow us to address a broader range of problems by using the Post K computer in new and innovative ways," says Matsuoka. "We see the Post K computer becoming an essential tool for solving problems in the areas of bioscience, disaster preparedness, energy security, environmental issues and manufacturing." ■



RIKEN Center for Computational Science (R-CCS)
Phone: +81 78 940 5555
E-mail: r-ccs-koho@ml.riken.jp
Web: www.r-ccs.riken.jp/en
Facebook: www.facebook.com/RIKEN.RCCS.en



The K computer



The RIKEN Center for Computational Science



Satoshi Matsuoka, director of the R-CCS

A visionary medical centre

THE NEW KOBE EYE CENTER combines cutting-edge disease research and comprehensive vision care

Ophthalmologists in Kobe have been trailblazers in using induced pluripotent stem (iPS) cells to treat degenerative eye diseases. But it was not until recently that they had a research hospital dedicated to eye and vision care to call their own.

The Kobe Eye Center opened its doors in December 2017. Gleaming with glass curtain walls that illuminate interior spaces — much as regenerative eye therapies restore light sensitivity to the retinas of people with macular degeneration — the new centre now serves as Japan's foremost hub for the advance of cutting-edge therapies for people with

intractable eye diseases and other visual impairments.

Situated on Kobe's Port Island, in the heart of the city's biomedical innovation cluster, the seven-floor facility includes a cell-processing centre, a research institute and a 30-bed ward, plus outpatient clinics for routine eye exams and other low-vision care services.

Having all these resources in one building will help scientists and doctors involved in the world's first-ever human studies of iPS cell-derived tissue transplants to accelerate research and better integrate new regenerative therapies with other social and behavioural aspects of patient

care, says Masayo Takahashi, head of the RIKEN Laboratory for Retinal Regeneration who helped lead construction of the ¥4 billion (approximately US\$36 million) centre.

Previously, these endeavours were spread across labs and clinical sites at the Kobe City Medical Center General Hospital, the Institute of Biomedical Research and Innovation, and the RIKEN Center for Biosystems Dynamics Research. "It's now easier to communicate among the groups by having them under the same roof," says Takahashi, who is leading the iPS studies, along with Yasuo Kurimoto, director of the Kobe Eye Center Hospital.

As well as pushing the boundaries of regenerative medicine and developing new kinds of diagnostic techniques, clinicians at the centre — which includes 11 full-time doctors, 12 part-time affiliates and another 12 research scientists — are using state-of-the-art technologies to promote rehabilitation for people with visual impairment. Patients can practise walking around an obstacle course to make the fullest use of their

remaining sight, and there's a driving simulator for people to assess whether it is still safe to get behind the wheel.

**IT'S
IMPERATIVE
WE CONNECT
MEDICAL
INTERVENTIONS
WITH
IMPROVING
THE DAILY
LIVES AND
WELFARE OF
OUR PATIENTS**

It is all part of an effort to create a hub for ophthalmological services in Japan, a country with an ageing demographic and where the rates of people expected to suffer from glaucoma, degenerative myopia, cataracts or other causes of visual impairment are set to soar within the next few decades. "It's imperative we connect medical interventions with improving the daily lives and welfare of our patients," says Takahashi. ■

 **Kobe Eye Center**

Kobe Eye Center

Tel: +81-78-381-9876

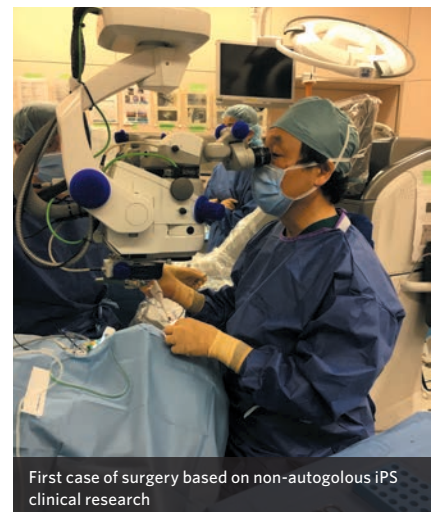
Web: www.kobeeyecenter.jp/english/



Kobe Eye Center at Port Island



Vision Park provides low-vision care services



First case of surgery based on non-autologous iPS clinical research

Kobe City Medical Center General Hospital

Providing vital help for everyone

KOBE CITY MEDICAL CENTER GENERAL HOSPITAL is strongly committed to providing emergency medical services for all

A simple philosophy underpins everything that happens at Kobe City Medical Center General Hospital — a hospital should be a place that those in urgent need can always rely on, no matter what the circumstances. “As Kobe’s flagship hospital, our duty is to be a place where people can seek medical aid even when others turn them away,” explains President Ryo Hosotani. “We place top priority in offering quality emergency medical services and advanced healthcare for the people of Kobe.”

With the rapid ageing of society in Japan, the number of patients requiring emergency medical services is on the rise.

Kobe City Medical Center General Hospital boasts an astonishing 99.1% take-in rate for emergency patients brought by ambulance — a remarkable achievement given the shortage of medical professionals in Japan. In 2017, as many as 35,000 people were admitted to the hospital’s emergency room, and the hospital has been ranked first in Japan for four consecutive years for emergency care in an assessment by the Japanese government.

While many hospitals turn away patients displaying mild symptoms, Kobe City Medical Center General Hospital makes a point of examining everyone, regardless of

how light their symptoms may appear. The hospital’s paramedics consider this approach to be crucial for identifying life-threatening symptoms before a condition becomes irreversible. “With over 40 years of experience or providing emergency care, we are extremely proud of our achievements,” says Hosotani.

OUR DUTY IS TO BE A PLACE WHERE PEOPLE CAN SEEK MEDICAL AID, EVEN WHEN OTHERS TURN THEM AWAY

The hospital is currently embarking on new initiatives to provide healthcare that meets the needs of all. “In Japan, there is an increasing emphasis on providing all needed healthcare within the local community. As a municipal hospital, we thus have a responsibility to offer non-profitable medical services or other services specified by the state,” explains Hosotani. “In addition, due to our merger with the Institute of Biomedical Research and Innovation

Hospital (IBRI) last year, we have a renewed mandate to spearhead clinical research.”

The hospital already leads the health industry in Japan in terms of the number of patients who receive state-of-the-art treatments, such as endovascular surgery for conditions affecting blood vessels and blood circulation in the brain, stem cell transplantation for blood conditions and personalized cancer treatment. In particular, it has the first hybrid operation room specifically for cerebrovascular treatment in Japan.

“The main driving force behind our hospital’s development has been exceptional teamwork,” says Hosotani. “Extraordinary power is unleashed when staff are motivated to work collectively towards a common goal.” ■



Kobe City Medical Center General Hospital
Tel: +81-78-302-4321
Web: chuo.kcho.jp/foreign_page/eng/eng_index.html



99.1% intake rate for emergency patients



Kobe’s flagship hospital for its citizens



President Hosotani, committed to delivering advanced health services

Amgen Foundation's move to scale up science education: Virtual laboratory experiences for students

In a lab, you can hypothesize. In a lab, you can run experiments. In a lab, you can learn to think—and operate—like a scientist. If you are a high school or college student and do not have access to a lab then you could be missing out on valuable experience—but not for long.

Thanks to a free virtual lab experience that will be made possible by the Amgen Foundation, headquartered in California, United States (US), and Harvard University, Massachusetts, US, students across the world will soon have the opportunity to tap into this important facet of science education online.



Students perform a science experiment in a classroom laboratory, which will soon be possible as a virtual experience.

The philanthropic arm of Amgen, the largest independent biotechnology company, is building a sophisticated portfolio of programmes that blend meaningful hands-on learning and the best of education technology to deliver on its commitment to reach more students with high-quality, curated and free science education programmes.

INTRODUCING LABXCHANGE—THE GAME-CHANGING NEXT STEP IN SCIENCE EDUCATION

On May 30 2018, the Amgen Foundation announced that it would contribute US\$6.5 million to Harvard University to create a free virtual lab experience and online community called LabXchange. The LabXchange platform, which will launch next year with a focus on biology, will also offer digital instruction and collaboration capabilities to high school and college students and instructors, enabling students to gain meaningful exposure to the scientific process.

Through LabXchange's virtual lab experience, students will be able to manipulate genes using plasmids. They will be able to

practice working with volumes of liquid smaller than one millionth of a litre. They will be taught how to produce thousands of copies of a specific sequence of deoxyribonucleic acid (DNA). There will be simulations ranging from engineering chimeric antigen receptor (CAR) T cells to attack and kill cancer cells to making macrophages resistant to human immunodeficiency virus (HIV) infection. There will also be a unit that demonstrates protein folding using simple experimental simulations, not to mention virtual experiments that mimic the differentiation from stem cells to beating heart cells in a tissue culture.

Soon it won't matter if students are living in Michigan or Malaysia. They will be able

to access an unparalleled lab learning experience—and do it for free.

"Advances in technology are not only having an incredible

**SOON IT WON'T
MATTER IF
STUDENTS ARE
LIVING IN MICHIGAN
OR MALAYSIA.
THEY WILL BE ABLE
TO ACCESS AN
UNPARALLELED
LAB LEARNING
EXPERIENCE—AND
DO IT FOR FREE**

impact on how we develop and deliver innovative medicines to patients, but also in how we educate and inspire the next

generation of scientists," says Robert A. Bradway, chairman and chief executive officer at Amgen. "By joining forces with Harvard, LabXchange's interactive educational platform will give students studying biology around the world access to a unique virtual lab experience for free, dramatically expanding the Amgen Foundation's reach in science education."

Giving students exposure to deeply immersive scientific learning experiences is not a new concept to the Amgen Foundation. In fact, LabXchange is just the latest addition to the Foundation's expansive portfolio of science education programmes that has steadily grown over nearly three decades. However, this

investment is significant in that it reflects a meaningful step towards scalable programmes that leverage the power of technology to open the world of science to young people everywhere.

"This unique virtual lab experience is designed to level the playing field for aspiring scientists globally while directly supporting and complementing our global science education portfolio," says David Reese, executive vice president of research and development and member of the Amgen Foundation board of directors. "As a career scientist, I have a deep appreciation for what this programme can accomplish by getting students to learn science by doing it."

For high school and college students, the acquisition of basic lab skills and early engagement in the scientific process can pave the way to more advanced scientific learning—and perhaps even a career in science—but sparking the curiosity of students today is not a given.

THE STORY OF STEM

There is no shortage of research on the skills and education gap that exists when it comes to so-called STEM (science, technology, engineering and maths) fields. One study, conducted by the United States-based Business-Higher Education Forum, shows that 80% of high school students are either not interested or not proficient in STEM subjects (www.bhef.com). Other findings suggest that the number of American students pursuing STEM careers is growing at less than 1% each year despite high demand for STEM jobs (www.act.org). Meanwhile, according to a study by The National Bureau of Economic Research, headquartered in

Massachusetts, US, children from high income families are ten times more likely to be inventors—a conclusion drawn by comparing the household income tax brackets to patent filers (www.equality-of-opportunity.org).

Needless to say, the challenges of science literacy and advancing a technically skilled workforce have broad implications across a host of industries, including healthcare and biotechnology and society as a whole. That is why the Amgen Foundation has made science education an area of focus for nearly three decades.

**80% OF STUDENTS
ARE NOT INTERESTED
OR NOT PROFICIENT
IN STEM SUBJECTS**
-BUSINESS-HIGHER
EDUCATION FORUM

THE AMGEN FOUNDATION'S COMMITMENT TO SCIENCE EDUCATION

The Amgen Foundation's first science education programme, developed through a collaboration between Amgen scientists and educators, is called the Amgen Biotech Experience. Over the course of approximately three weeks and under the guidance of their science teachers, student participants produce a recombinant DNA molecule and use it to transform the bacterium *Escherichia coli*. Breaking from the traditional textbook-driven classroom format, students better understand how science is used to develop medicines and are able to imagine themselves as scientists. What started as a single class in 1991 has developed into an innovative programme with more than 700,000 student participants from schools across the US, Canada, Europe, Asia and

HOW DO YOU MEASURE THE IMPACT OF A PHILANTHROPIC INVESTMENT? RUN THE EXPERIMENT.

As the philanthropic arm of a company that relies on data to make some of its biggest decisions, the Amgen Foundation recently decided to conduct an evaluation to test the effectiveness of one of its flagship programmes. The Amgen Biotech Experience puts 80,000 students per year in the shoes of a real-life biotechnology company scientist. But the Foundation wanted to verify with data that this programme does in fact help them to deliver on their commitment to science education.

"We had anecdotal data showing the impact of the programme for many years, with high school science teachers and students raving about the ability to transform a living cell into a protein factory," says Eduardo Cetlin, president of the Amgen Foundation. "However, with the programme coming into a new funding cycle in 2017, which would include an expansion to nine new international locations, we felt it was critical to ask: Is the programme actually making a difference when it comes to inspiring and educating students in science and biotechnology?"

To find the answer, the Foundation called upon WestEd, a notable independent education research organization headquartered in San Francisco, California, US, to conduct a comprehensive evaluation of the programme. By asking 3,500 high school students a series of questions before and after they went through the Amgen Biotech Experience, WestEd found that students had: (i) significant and substantial learning of biotechnology and (ii) increased interest and confidence in doing science and biotechnology. Students showed a statistically significant increase ($p < 0.001$) and with large effect size ($d = 1.03$) on a 25-question validated assessment, with an average increase of 20 percentage points between the pre- and post-tests.

"It was incredibly rewarding to see the data validating what we knew in our hearts," notes Cetlin. "Now with LabXchange, we hope to augment the programme offerings to incorporate cutting-edge science experiments that will further enhance the student learning experience and allow them to see the promise of biotechnology firsthand."

Australia. Students in the Amgen Biotech Experience get hands-on experience with cutting-edge biotechnology tools and perform wet-lab procedures.

The next programme the Foundation established is the Amgen Scholars Program. Launched in 2006 and sustained by a 12-year, \$50 million commitment from the Amgen Foundation, the Amgen Scholars Program invites qualified undergraduate students from hundreds of

colleges and universities to conduct groundbreaking research at the world's leading institutions under the mentorship of world-renowned scientists. The programme is designed to allow exemplary students from all economic backgrounds to participate. More than 3,900 students from 700 colleges and universities have completed the programme, with the vast majority now pursuing advanced degrees and careers in scientific fields. In 2017, Amgen Scholars



Inside the Labs of Amgen

Some of Amgen's most important innovations have taken place inside its very own labs. Here are 5 legendary Amgen scientists who helped change the practice of medicine.

Amgen scientists pictured clockwise, from top left: Fu-Kuen Lin cloned the erythropoietin gene leading to Amgen's first medicine; Larry Souza cloned the G-CSF gene leading to Amgen's first major oncology medicine; Simon Jackson helped to elucidate the biology of the PCSK9 protein leading to Amgen's first cardiovascular therapy; Cen Xu advanced an antibody approach against the CGRP receptor that led to Amgen's first migraine medicine; David Lacey helped discover the osteoprotegerin (OPG) protein leading to Amgen's first osteoporosis medicine.

alumni published 712 works and received 810 awards, including one named to Forbes's '30 Under 30' in healthcare.

EMBRACING ONLINE SCIENCE EDUCATION TO EXPAND REACH

More recently, the Amgen Foundation has been focused on expanding the reach of its science education portfolio using technology, recognizing that the educational landscape is becoming increasingly technology-enabled. For example, according to one survey of teachers whose pupils range from pre-school to 18 years old, one in three teachers are using technology to encourage working in teams and collaboration, 58% are using tech tools for project-based learning and 55% of teachers are using technology to encourage creative thinking (www.kahoot.com).

The first move in this direction is collaboration with the Khan Academy. The Khan Academy is a free online learning platform, making high-quality, video-based educational content accessible to everyone. In 2017, the Amgen Foundation became the exclusive

sponsor of the Khan Academy's biology content, supporting the programme with a three-year, US\$3 million grant.

Soon after, came discussions that led to the investment in LabXchange. The Amgen Foundation explored for some time how best to connect and virtually scale its high school lab and undergraduate research initiatives. This exploration ultimately led to Robert Lue, a Harvard University professor of molecular and cellular biology, who brings deep expertise in online learning, science education, and the engagement of students in lab and research experiences.

Following the US\$6.5 million contribution from the Amgen Foundation, the team at Harvard University is building prototypes and testing them with potential users in 2018, with plans to launch globally in 2019 with a focus on biology.

"There are many millions of students who, as a result of economic or geographic limitations, simply do not have access to one of the most central aspects of being a scientist, which is working in a laboratory," says

Professor Lue. "LabXchange addresses this issue with a platform that integrates dynamic experimental simulations with background curriculum and social networking—all created to more effectively expose students of varying backgrounds to the authentic and engaging experience of scientific discovery."

Students want to learn wherever and whenever it is convenient for them and platforms like LabXchange have the potential to deliver the goods to students of all incomes and backgrounds.

INSIDE AMGEN'S LABS

When you think about what sort of programmes the Amgen Foundation might support, it's no great leap to imagine they might be related to science or, for that matter, experiences that might immerse you in a lab. After all, over the course of more than three and a half decades, Amgen

went from a small, venture-backed start-up with a handful of scientists operating out of a small building in an unassuming strip mall in southern California to a company that employs many thousands of scientists around the world.

How did that happen? One could argue that it was largely because of what scientists were accomplishing in Amgen's own labs. After all, scientists working in labs at Amgen have made discoveries that led to innovative and meaningful medicines for kidney disease, cancer, osteoporosis, cardiovascular disease and migraine.

As a result, the enthusiasm at Amgen from staff, senior scientists and senior leadership to support the advancement of aspiring scientists through the use of a lab experience should come as no surprise. And this enthusiasm can be seen through the Foundation's long-standing commitment to science education.

"We think the school science lab is a magical place—a place where, as a young student, you get to do what you read about in a science book," says Cetlin. "As students move through their educational journey, the lab is the perfect setting for them to start thinking like scientists: asking questions, crafting hypotheses, designing and running experiments, failing and starting the cycle again. That's why we have invested tens of millions of dollars to give students the chance to experience lab learning firsthand."

For more information about the Amgen Foundation, visit www.AmgenInspires.com and follow us on Twitter @AmgenFoundation. For more information about LabXchange, visit www.LabXchange.org and follow @LabXchange on Twitter.

CERISE SKY MEMORIES

A gift to remember.

BY WENDY NIKEL

I remember a childhood that didn't exist. Hot apple pies cooling on park benches. Small toes pressed into scorching white sand. Snowball fights leaving crisp, crunchy ice crusted in the collar of my coat.

A tangle of neurological connections to construct a lifetime that never was. Things that never happened. Places that don't even exist.

We whisper about these memories sometimes, Marina and me. She's the only other X3-Model left in the office. The only other employee-asset remaining from that brief phase in biotechnological progress before designers began second-guessing the cost of programming us with such concocted complexity.

"It cost ten thousand dollars per memory," Marina muses as she swirls her spoon around her yogurt. It's strawberry, as usual. She buys it by the gallon and brings it to work in dainty little Tupperware bowls that she stacks beside the white cube food-packs that the rest of us apathetically consume. I've always suspected it has to do with her childhood, but I don't ask — not in front of the others. No need to draw more attention to the fabricated pasts that they view as defects, signs of our antiquity.

"If you could buy one more," she asks, "what would you pick?"

"I don't know." I flatten my empty food-pack and keep my voice low, head down. "We ought to get back to work."

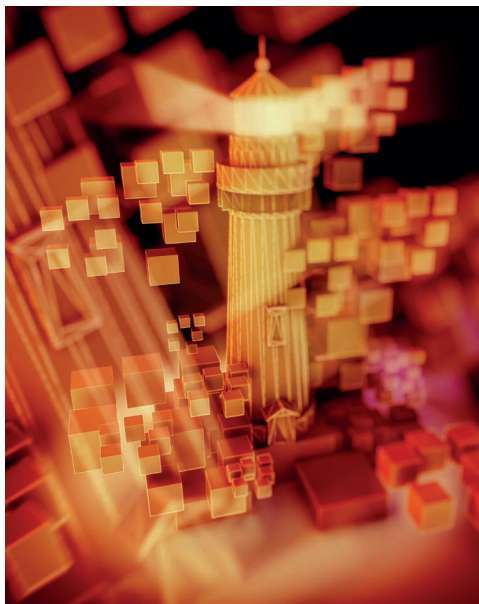
"I'd buy a birthday party," she says, staring at her spoon. "One with family. A mother. Father. Siblings."

"They wouldn't be real." Marina and I are the closest we've got to siblings — members of the same genetic batch, products of the same vats, designed for proficiency in the mind-numbing tasks of corporate sorting and filing.

Marina shrugs and licks the spoon. "At least I'd have someone to remember."

In my favourite memory, I'm sitting on a front-porch swing, my legs tucked beneath me and a book on my lap, watching a cerise sunset over a vast, windmill-dotted field. A breeze blows across my face, and I close my eyes, breathing in rich loam.

Marina says that's why I like the quiet — why city life sets me on edge. A glitch, the company would call it if they knew, which is why I work so hard to keep from startling when the boss calls out my number, summoning me to his office.



I've never been in his office before.

It's a small room with buzzing, yellowed light. A trio of framed pictures sits upon a desk too large for the space, and one of them depicts a scene so familiar, I can't help but stare, can barely comprehend his words:

"You're being decommissioned."

"Decommissioned?" I stare not at him, but at the lighthouse in the photograph. The cliff. The waves. I can almost feel the sand in my toes, taste the salt on my lips.

"We're bringing in new X14-Models," he continues. "When you punch out today, we'll remove your chip and you'll be free to pursue other employment, no longer company responsibility."

No longer company responsibility, meaning they will no longer provide me with food, shelter, clothes. *Decommissioned*, meaning declared obsolete.

"Any questions?"

"Yes." I point to the picture. "Where was that taken?"

Before I punch out, I slip Marina a note that I hope makes her own decommissioning easier: *They're real. I've seen it. The places from our childhood exist.*

I buy a bus ticket I can't afford to a state where I've never been and hope that whoever programmed that lighthouse in my mind was drawing from their own experiences, and that the farmhouse will be nearby.

My toes press into the scorching white sand. I lick my salt-cracked lips.

A hot-dog vendor lends me his pen, and I scribble a napkin-sized sketch, but none of the beachcombers recognize the windmill field. None knows of jobs for decommissioned X3s.

The sand turns gritty. My shoulders burn. The sun dips low towards the horizon, and doubts creep in with the cold.

I shouldn't have left the city. I shouldn't have assumed that because one memory was based on fact, all of them must be real.

Yet it doesn't stop me from asking two ... four ... ten more times. It doesn't stop my tears when someone finally says, "I know where that is."

I stand on the doorstep and wring my hands, unsure what I hope for, yet hoping against hope, and when a woman in overalls and a red bandana answers, she looks just as confused as I feel.

"Do you live here?" I ask. "That is ... have you lived here long?"

"Grew up here. You're an X3-Model."

"Yes."

"I worked on your kind, back in the day."

"In memory modifications?"

She raises a greying eyebrow. "How'd you know?"

I want to ask about the porch swing, about the book, about the sunset itself — things so important she infused them into her work, leaving a sliver of herself within me.

I want to ask how she ended up here — if, when the memory mods were discontinued, she became obsolete, too. *No longer company responsibility.*

I want to ask if she needs anything sorted or filed. I want to ask about the dirt on her knees. I want to ask about snowball fights and apple pie and whether the city noise makes her nervous, too, and if that's why she returned to this quiet place with gentle breezes that smell of loam.

But she's the one to make the first move, to nudge the screen door open. "Why don't you join me on the porch for some tea? It seems we might have some things in common."

I follow, in steps of hope, as a cerise sunset lights the sky. ■

Wendy Nikel is a speculative fiction author with a degree in elementary education, a fondness for road trips and a terrible habit of forgetting where she's left her cup of tea. For more info, visit wendynikel.com.

ILLUSTRATION BY JACEY

natureOUTLOOK

SCIENCE AND TECHNOLOGY EDUCATION

4 October 2018 / Vol 562 / Issue No 7725



Cover art: Sébastien Thibault

Editorial

Herb Brody,
Richard Hodson,
Brian Owens,
Elizabeth Batty,
Nick Haines

Art & Design

Mohamed Ashour,
Wesley Fernandes,
Andrea Duffy,
Denis Mallet

Production

Nick Bruni, Ian Pope,
Karl Smart

Sponsorship

Samia Burridge,
Anushree Roy

Marketing

Nicole Jackson

Project Manager

Rebecca Jones

Creative Director

Wojtek Urbanek

Publisher

Richard Hughes

Editorial Director

Stephen Pincock

Magazine Editor

Helen Pearson

Editor-in-Chief

Magdalena Skipper

A strong background in science, technology, engineering and mathematics (STEM) is vital for more than budding scientists. Future jobs in a wide variety of areas will require skills in STEM subjects. This Outlook explores how science education is being modernized to prepare students for life in the twenty-first century.

The way science teachers are being trained is evolving to take account of new knowledge about how students learn. Trainee teachers can now use 'design thinking' to create active, hands-on lessons that keep students engaged (see page S2).

Some science lessons now use virtual labs from companies such as Labster. These give students the freedom to experiment without the safety and financial constraints of the real world. But could virtual labs ever replace the real thing (S5)?

Art is often seen as separate from science, but drawing can be a valuable way to deepen students' understanding of a subject. There are ways to bring drawing into science lessons even for those who think they can't draw (S8).

Science education is vital for the developing world as countries strive to modernize their economies and improve conditions for their citizens. Bringing innovations from the developed world, and adapting them to the local context, can help to close the gap between rich and poor nations (S10).

The lack of diversity in science must be addressed at all levels of education. New initiatives in physics are helping to increase the proportion of women and scientists of colour, and may hold lessons for the rest of science (S12).

Technology is being integrated into education at earlier ages, from the use of tablets in the classroom to lessons in coding. A digital-intelligence project aims to help students learn the skills they need to be safe online (S15).

We are pleased to acknowledge the financial support of the Amgen Foundation in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Brian Owens

Contributing editor

CONTENTS

S2 TRAINING

Building better science teachers

Improved teacher training can help students learn

S5 ON-SCREEN LEARNING

The virtual lab

Can simulations match the experience of a real lab?

S8 PERSPECTIVE

Drawn to science

Bethann Garraon Merkle shows how art can improve understanding

S10 DEVELOPING WORLD

Expanding the reach of science

The transfer of ideas can help developing countries

S12 SOCIAL POLICY

Drive for diversity

Efforts to attract physics graduates from underrepresented groups

S15 DIGITAL EDUCATION

The need for digital intelligence

Children need help to stay safe online

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2018).

VISIT THE OUTLOOK ONLINE

The *Nature Outlook Science and technology education* supplement can be found at www.nature.com/collections/science-education-outlook. It features all newly commissioned content as well as a selection of relevant previously published material that is made

freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, <http://nature.asia/jp-contact>; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, <http://nature.asia/jp-contact>; China, <http://nature.asia/china-subscribe>; Korea, www.natureasia.com/ko-kr/subscribe.

CUSTOMER SERVICES

Feedback@nature.com

Copyright © 2018 Springer Nature Limited. All rights reserved.



TRAINING

Building better science teachers

The latest training techniques emphasize classroom practice and design thinking.

BY JOSHUA HATCH

If you spend time with a young child it soon becomes clear that the astronomer Carl Sagan was right when he said: “Every kid starts out as a natural-born scientist...”. Spend time with a typical high-school student and it’s clear that the second part of Sagan’s quote is also correct: “...then we beat it out of them. A few trickle through the system with their wonder and enthusiasm for science intact.”

Attempts to change that dynamic are increasingly focused on teachers, particularly the way they are trained and how they interact with students. In the United States, the Next Generation Science Standards, developed by the National Research Council, outline ways for teachers to encourage student enquiry, feed their curiosity, and deepen their understanding of scientific concepts. Meanwhile, the latest

teacher-training techniques place a greater emphasis on pedagogy and classroom practice as another way to improve science, technology, engineering and mathematics (STEM) education. Both approaches are seen as models for other countries that seek to improve their own STEM education.

Existing teaching methods have long been based on “the rhetoric of well-established conclusions”, according to Jonathan Osborne, professor of science education at Stanford University in California. “The dominant paradigm that most teachers work with,” Osborne says, “is essentially: ‘I know and you don’t know, and I’m here to communicate it to you and explain it to you.’ And the problem with that is we know it doesn’t work very well.”

Not only does that approach turn off students but it may also be failing society. A 2013 report¹ from the US National Science and Technology

Council stated that “current educational pathways are not leading to a sufficiently large and well-trained STEM workforce.” The report further blamed the educational system for failing to produce a STEM-literate public. As a result, up to half of those who want to pursue a STEM education in college are ill-prepared by their secondary schools, according to a report² by the educational-testing organization ACT. Yet the demand for STEM graduates remains high. The European Commission set a goal in 2011 of adding 1 million science researchers by 2020, and in 2012 then-President Barack Obama set a target of 1 million new STEM graduates by 2025. In response, many secondary schools have been increasing their maths and science requirements.

The number of high-school students enrolled in maths and science courses rose by more than 60% in the 20 years from the mid-1980s in the

SEBASTIEN THIBAUT

United States, according to a study³ from the University of Pennsylvania in Philadelphia. But rises do not directly improve STEM education — they just add more pressure to the system.

Improving outcomes, says Peter McLaren, executive director of the non-profit initiative Next Gen Education, will require a shift in the classroom from a teacher-centred approach to one that helps students work through concepts themselves. McLaren, who helped to write the Next Generation Science Standards, used to teach general science in East Greenwich, Rhode Island, and recalls a lesson he once taught on gases. He had borrowed his ex-wife's perfume bottle and sprayed it around the front of the classroom. "I asked the students to raise their hands when they could smell something," he recounts. The kids were engaged and excited, and then McLaren proceeded to explain what was happening to the gas molecules. The kids' enthusiasm was quickly snuffed out by his 'let me explain' lecture.

What he should have done, he now knows, is ask his students: "What would cause the scent of the perfume to reach all the way to the back of the classroom?" and start a discussion. Asking that question would empower the students to use their knowledge and imagination to develop scientific ideas about the concept being discussed. Instead of 'learning about,' McLaren says, "it's about 'figuring it out'."

BETTER BY DESIGN

Development of the Next Generation Science Standards has been a multi-year, state-led effort, based in part on standards in ten countries including the United Kingdom, Finland and Japan. At the same time, some of those same countries are looking to the United States as a model for how STEM education can be improved. Meredith Portsmore, director of the Center for Engineering Education and Outreach at Tufts University in Massachusetts, says that educators from around the world have enquired about the US approach.

"They are starting to see some of the same issues, with students not finding STEM appealing because it's not creative," Portsmore says. The educators see integrated learning — like that outlined in the Next Generation Science Standards — as a way to create more innovative and creative scientists and engineers.

However, some of the concepts might not translate easily. Some systems are built around strict standards or large class sizes that would make it difficult for teachers to give students the focused, personalized guidance they need.

The Next Generation Science Standards are supported by large organizations such as the National Research Council, the National Science Teachers Association, and the American Association for the Advancement of Science, but smaller entities are also trying to change the way science is taught in the classroom.

The Woodrow Wilson National Fellowship Foundation in Princeton, New Jersey, was founded in 1945 to address a shortage of



Woodrow Wilson Academy design fellows prepare a hands-on science lesson for local students.

college faculty following the Second World War. Since 2008 it has been collaborating with US universities to revamp teacher education. Deborah Sachs, director of the University of Indianapolis' Teach (STEM)³ programme, says that in one such partnership the fellowship worked with the University of Indianapolis in Indiana to answer the basic question: "What is it that effective STEM teachers need to know in order to be successful?" The idea was to rethink how teachers were trained from the ground up.

One realization was that trying to explain concepts and then have students apply them — or worse, simply regurgitate them — did not work. Instead, teachers should create projects in which concepts become apparent as students work through real-world challenges.

Sachs cites an example in which a geometry teacher working on a lesson about circle circumference and diameter had students map phone towers and their signal strengths. As students worked through the project, they saw for themselves where there were gaps in mobile-phone coverage and developed a deep understanding of the mathematical concepts.

Since its start in Indiana, the foundation has expanded its teaching programme to 5 other states, including 31 universities. But now the foundation is starting its own teacher-training academy using one of the technology industry's favourite buzz phrases: design thinking.

The term 'design thinking' originated in the 1960s with the idea that design should be rational and solve problems, rather than simply focusing on aesthetics. It involves solving problems through research, empathy, ideation and prototyping, and was famously used in the 1980s to create Apple's first computer mouse. In that instance, the goal was not to design a snazzy new pointing device, but to give non-technical people the ability to easily

use a computer. Since then, design thinking has been used to create thick-handled toothbrushes for children — it is easier for their small hands to grip a big handle than a small one — and to launch disruptive companies such as Airbnb.

MAKING A CONNECTION

Located in a small office just a 15-minute walk from the Massachusetts Institute of Technology in Cambridge, the Woodrow Wilson Academy of Teaching and Learning is home to a handful of staff and incoming students. They are busily identifying some of the problems with current STEM teacher training and are working on ways to overcome them. One challenge is to give teachers more practice managing classrooms, working with colleagues and even dealing with parents.

The staff and students came up with several ideas to address this problem. One idea was to send student teachers to Boston's Museum of Science after-school clubhouse, where they work with kids on various projects including 3D printing, video editing and website development. The students do not have to be there, so teachers need to connect with them and keep them engaged if they want the students to stick around.

Doyung Lee helped to set up the academy's curriculum in 2017 as a design fellow and will attend as a student in autumn. He says that volunteering at the clubhouse taught him how to create a welcoming and safe environment, partly by being curious about his students' entire lives. "You can be great with content knowledge or pedagogy," says Dan Coleman, the academy's chief learning and design officer, "but if you can't connect with students, you're going to fail."

Lee's experience at the science museum also



Science teachers benefit from practising experiments before performing them with students.

taught him how classrooms can suppress a student's enthusiasm, rather than tap into it. He recalls a middle-school student who came to the clubhouse and loved video editing and animation. The student was eager to expand his skills and wanted to share what he was learning, but he was struggling at school.

This is exactly the sort of situation that the Next Generation Science Standards seek to address, says McLaren. "When kids go to an informal education opportunity," he says, such as the science museum's clubhouse, they get to pursue their interests and that excites them. But that's not what typically happens at school. "In the classroom, we say, 'that's very nice, but we're doing something else today,'" says McLaren, and that squashes their eagerness. It would be better, he adds, to connect a lesson to existing student interests and then guide them within the constraints of the curriculum.

Imagine a student with an interest in video production taking a class on environmental science. By asking the student to make a video documenting an environmental-science process, such as the water cycle, the student's excitement for making the video carries over to the scientific enquiry. "If you give that kind of freedom to students," McLaren says, that's when the magic happens.

PRACTICE MAKES PERFECT

Another idea being tried at the academy is to have students practise real-world situations through a computer simulator. For example, teachers often have to deal with upset parents but rarely get to practise that encounter before it happens in real life. The simulator lets the student teacher try out different strategies and find out what works and what doesn't. "People come through the experience sweating," Coleman says. "It feels like a real encounter."

Gaining such experience as student teachers is critical because a lack of classroom experience is inversely related to teacher effectiveness — and directly related to teacher turnover. "We found huge variations in how much practice in student teaching prospective teachers get," says Richard Ingersoll, a sociologist at the University of Pennsylvania's Graduate School of Education. "Something like a fifth of all new hires have never had practice. Their first day of teaching is their first day with kids. We found this was especially true for new science teachers."

A lack of practice as a student teacher translates to a lot of science teachers leaving the profession. Although this is applicable to teachers of any discipline, "it is particularly dreadful for science teachers, because they are the group that most frequently hasn't had any student teaching," Ingersoll says.

Although the Woodrow Wilson academy has yet to graduate any students of its own — its first class of 21 future teachers began in September — more than 80% of teachers from the foundation's work with existing teaching colleges, such as the University of Indiana, have remained in the profession five years after entering. This figure is similar to the national average, according to a 2015 study⁴ by the US Department of Education.

Because STEM often relies on technology and experiments, it is important to practise its use in the classroom. One of the academy's faculty members, Andrew Wild, knows this first-hand. He earned a PhD in science education from Stanford University in California before working as a science teacher in the San Francisco Bay Area. One day he went to his

classroom of 42 students prepared to teach a lesson on circuits. But it didn't go according to plan. "I remember the wires going across the room," Wild recalls, "and students were tripping on wires, pulling them from the circuit boards." There weren't enough outlets and the lesson was a near disaster. Although Wild was well-versed in the subject matter, his failure to consider space challenges and to prepare for technical problems undermined the lesson.

MODEL TEACHERS

Another important point in teacher training, McLaren says, is to teach the teachers using the methods you want them to use in the classroom. This is referred to as "modelling the model" by the academy's vice-president for strategic initiatives, Deborah Hirsch. "In order to teach differently, you need to learn differently," she says.

McLaren agrees and describes an example from a professional-development workshop. "The teachers have to be put in the position where they are playing the role of the student," he says. So to convey one of the Next Generation Science Standards such as understanding causation, teachers would be presented with classroom situations and asked to identify causes and patterns. As a result, the teachers are using evidence based on their investigation to arrive at an explanation. The 'aha!' moment, he says, is when "teachers can see how a shift in their instruction makes a world of difference."

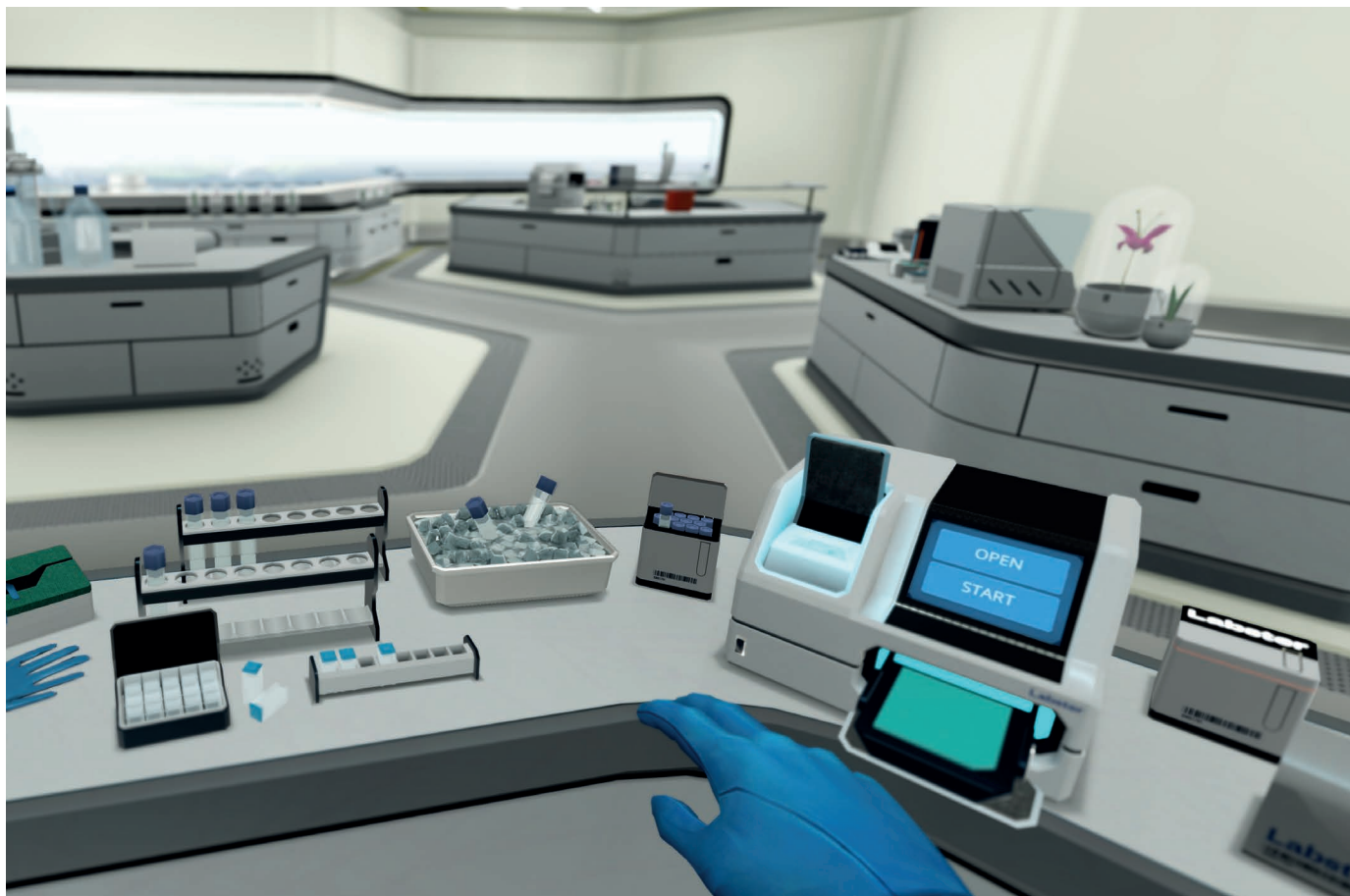
Osborne also encourages a shift in the way STEM teachers engage with their students. He advocates teaching argumentation in science as a way for students to understand scientific concepts. "The history of science is the history of vision and argument," he says, noting long-standing debates about the workings of the Solar System and the cause of disease. This is not an easy assignment for teachers, he admits: "It's a relatively complex skill."

McLaren says it takes time for teachers to learn these skills, and he was no exception. In 2001, he won the Milken Educator Award, which is given each year to exemplary early-to mid-career teachers in each state. Even so, looking back on his teaching career, and considering how the Next Generation Science Standards shift the focus away from teachers and towards students, he says: "I want to write a letter of apology."

What would he write? "I didn't give you enough opportunity to choose your path of investigation... I should have let go more and have you figure it out more." ■

Joshua Hatch is assistant managing editor at The Chronicle of Higher Education.

1. Federal Science, Technology, Engineering, and Mathematics (STEM) Education: 5-Year Strategic Plan (National Science and Technology Council, 2013).
2. The Condition of College & Career Readiness 2016 (ACT, 2016).
3. Ingersoll, R. *Kappan Mag.* **92**(6), 37–41 (2011).
4. Gray, L. & Taie, S. *Public School Teacher Attrition and Mobility in the First Five Years* (NCES 2015-337) (US Department of Education/National Center for Education Statistics, 2015).



Labster's enzyme-kinetics simulation allows students to feel as if they are in a real laboratory.

ON-SCREEN LEARNING

The virtual lab

Can a simulated laboratory experience provide the same benefits for students as access to a real-world lab?

BY NICOLA JONES

When I enter the lab, I see an open flame on an unattended Bunsen burner. The fume hood is open and a pile of explosive chemicals sits in the middle of the floor. In a fit of devilish abandon, I take off my lab goggles and pour an unknown liquid into a dirty beaker. It explodes, spraying acid into my eyes.

Don't worry, this is only a computer simulation, one of 70 produced by the Danish company Labster, based in Copenhagen. The experience looks and feels like a video game but its purpose is more serious: supplementing, or even replacing, laboratories for students who are unable to afford or access the real thing.

Arizona State University (ASU) has launched its first fully online biology degree course that uses simulations instead of actual lab work.

Labster has collaborated with Google Daydream to provide 30 three-dimensional (3D) lab simulations for the course, and says that more universities are likely to follow, including the University of Texas at San Antonio.

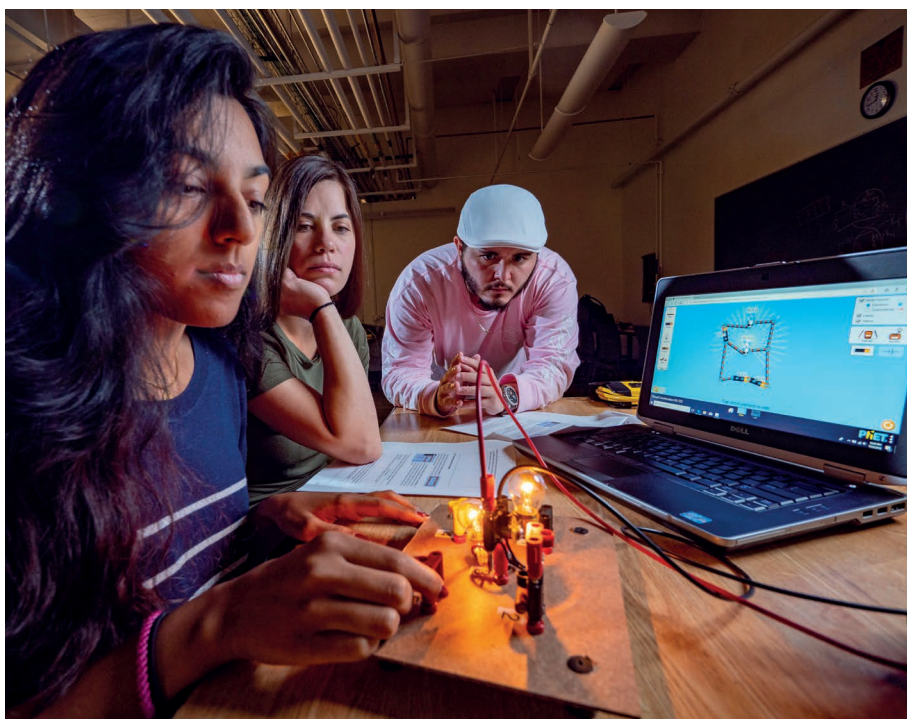
The idea of 'virtual labs' is gaining traction as companies and institutions try to expand their reach, cut costs, enhance student understanding, and provide a different kind of hands-on training for future scientists. For example, edX — the world's largest non-profit platform for free, online university courses — is due to launch its own lab simulations later this year in a project called LabXchange.

"The real lab is actually very limiting," says Brian Woodfield, a physical chemist at Brigham Young University in Provo, Utah, who has been developing virtual labs for decades. "A lot of it is toxic. We can't let them blow up things. We have limited time, and there are safety issues all over the place." The virtual

world, by contrast, is free of such restrictions.

Virtual labs range from stripped-down 2D video games, which use simple toggles to control a handful of variables, to 3D simulations that aim to provide a more immersive experience. Some provide students with an objective and step-by-step instructions, guiding them through the technical steps for carrying out complex procedures. Others are completely open ended. Woodfield's Virtual ChemLab, for example, lets students choose chemicals from a store-room shelf and use them however they want. "That's real chemistry," he says.

The inventors of virtual labs extol the benefits of technology for improving education, and emerging data suggest they are right: virtual labs do improve some test scores and help students to prepare for real-life scientific investigations. But there is still debate about whether they really enhance learning over what a textbook can provide; whether fancy



Students at the University of Colorado Boulder use PhET simulations alongside real lab equipment.

graphics are helpful or just a flashy distraction; and whether simulations really can (or should) replace real lab experience.

STORY BOARD

Labster and other similar products, including Late Nite Labs from Macmillan Learning, try to make users feel as if they are really in a lab. Labster's simulations have shadow and light, and allow users to 'walk around' — I spent much of my first few minutes in their sim orienting myself, working out where the doors were and how to navigate the space. A robotic voiceover guides users through the tasks, from putting on a lab coat to operating a DNA sequencer. When a procedure involves physical acts, they have to be mimicked by the user: a pipette must be picked up and a tip added, and then used and thrown away, each step performed with the click of a mouse or trackpad. The interface can be clunky and frustrating at times, and the required acts are repetitive, which is arguably a fair simulation of reality.

The virtual lab world is clearly not backed with the same dollars as the video-gaming industry. But there is still attention to detail: the labs are populated with characters who talk to you, and there is a view out of the windows; there are even quirky wall posters. Such details may not be necessary for teaching, but they help to enhance the immersive quality. The theory is that users who feel they are really in a space will devote more time and attention to it, and so benefit more from the experience.

Labster's stock of simulations ranges from an entertaining romp through basic lab safety, such as the 2D sim that sprayed my eyes with acid, to highly technical procedures including cancer sample preparation for mass

spectrometry. The price ranges from US\$10 for one simulation to \$199 for full access per student per term, although this is under review. Most simulations are in the fields of biochemistry and medicine (including viral gene therapy and DNA sequencing). These topics, says Labster's co-founder Michael Bodekaer, tend to be "too dangerous, too time consuming or too expensive" to do in a real laboratory.

Many of their simulations use storytelling to make them more engaging. In the animal-genetics lab, for example, the player visits a farm to sample meat, before learning how to develop a DNA test for double-musled cattle (a genetic variation that allows animals to have more muscle fibre and less fat) and then playing detective to find out whether meats labelled as organic are abiding by the rules.

There are clear benefits to this 'gamification': it seems to increase student motivation, which is not a trivial achievement. But whether any one simulation actually improves understanding or just lets students have fun blowing things up depends on the fine details of how they are designed.

"There has to be support for comparing your prediction to what happened. You have to be able to go back and forth, or run the simulation under different circumstances," says Marcia Linn, an educational psychologist at the University of California, Berkeley, who has been looking at virtual labs since the early days of Apple computers in the 1980s. Students also need a chance to summarize and reflect on what they have learnt, Linn adds. "If students don't have an opportunity to reflect, it's pretty common for that not to be effective."

Simulations can be particularly useful, she says, if they let students play around with

things that are normally outside their control. "The ones that are really valuable are the ones that allow students to explore complex phenomena that you can't explore with the naked eye: rapid airbag deployment, chemical reactions and climate change," says Linn. "We couldn't do hands-on investigations of those before." The graphics do not need to be fancy: you just need to see circles representing molecules dancing around to illustrate temperature or pressure, for example, or lines showing weather fronts moving around.

At Berkeley, says Linn, the science faculty often use PhET Interactive Simulations in their courses — free online sims produced at the University of Colorado Boulder that explore core science concepts, such as how a circuit works or what controls pH. PhET was founded by physicist Carl Wieman in 2002, a year after he won a Nobel Prize for his work on Bose-Einstein condensates, as he turned his attention to science education. PhET sims are now used about 90 million times a year.

These sims have simple graphics and allow a large degree of freedom for action within the confines of exploring a simple scientific principle. Users might, for example, make waves in a wave tank and watch them interact. Students can set whether the waves are sound, light or water; where the waves originate; their frequency and amplitude; and whether they go through a slit at the far end of the tank. In this way, the basics of wave physics are played out in an easily tweakable demonstration. The graphics are often no flashier than those in a typical textbook, but the interaction aims to provide more intuitive 'aha!' moments than you would get from a static reading exercise.

"Ours do not look and feel like an actual lab," says Kathy Perkins, director of the PhET Interactive Simulations project at Boulder. But that's not the point: they focus instead on optimizing understanding by giving students a lightly guided system to explore. "Labster is very storyboarded. Others provide a huge degree of freedom. We try to hit the sweet spot."

HIGH SCORE

Many studies have found little difference in learning outcomes between students who do virtual lab experiments and those who do them for real, whether it is undergraduates learning about heat exchange, or children at elementary school investigating springs¹. The main difference is that you cannot physically touch anything in a virtual lab, but this has surprisingly few limitations. The studies that show a detriment typically involve a completely unfamiliar physical task, such as children aged 5–6 learning to use a balance beam.

A few quantitative, controlled studies of more-immersive simulations are also starting to emerge. Mads Bonde, Labster's co-founder and chief executive, worked on one of these at the Technical University of Denmark. Bonde and colleagues gave half of the students in a first-year life-sciences class access to a Labster

simulation of a crime-scene investigation. When tested, the students who did the simulation scored, on average, 76% higher than those exposed to traditional teaching² (the difference disappeared when the students swapped groups in the second half of the study). Perhaps unsurprisingly, 97% said they felt the simulation made the course more interesting.

Kambiz Hamadani, a biochemist at California State University San Marcos, has been using Labster in his classes for two years. “Funding is tight, and space is limited, and we have a lot of students of different sorts and we have to teach them all,” he says. Virtual labs are well suited to those situations, he explains, allowing students to work at different rates from home. As well as using virtual labs to improve understanding, Hamadani also uses them to cope with the limited space in real labs, avoiding scheduling conflicts by shifting some lab studies into the virtual realm.

Hamadani secured a grant through the California State University system to redesign his courses using Labster and assess the shift. In 2016, he essentially copied Bondé’s procedure, using a Labster simulation of enzyme kinetics. He had a small class of 45 students, and the test involved only a few dozen questions, but Hamadani still saw improved results. On test questions that delved into higher-level understanding by requiring the application of learned ideas, those with access to Labster did 40–50% better than their peers. However, the recall of facts and definitions took a hit. “The Labster students are clearly diverted from textbook learning — their performance on textbook-type questions actually drops,” Hamadani says.

The students were enthusiastic about it, he says, and that’s important. “At the end, they said things like ‘that’s awesome’. Some of them came to ask for access to other labs that I didn’t even assign.” But other students felt overburdened if given more than a couple of virtual labs. In his second year of testing, he concedes, he “went a little overboard” by using up to five Labster simulations. “I overloaded them with too much work — they weren’t able to focus.”

VIRTUAL OVERLOAD

Hamadani used the 2D version of Labster. He says the 3D one would have been both too expensive (it requires more hardware, including smartphones and viewers) and potentially not worthwhile from a learning perspective.

High levels of immersion through 3D technologies are thought to increase engagement and emotional investment. In theory, this should set the stage for more in-depth learning, says Guido Makransky, an educational psychologist at the University of Copenhagen who puts virtual labs through their paces. But spicing up a potentially boring topic with an immersive 3D experience does not always help, he adds, because students might get distracted. “Learners are curious so they try to play with things to see what will happen,” says

Makransky. They spend their time wandering around the virtual world, rather than focusing on the task at hand.

Makransky let 52 university students play either the 2D or 3D version of one of Labster’s simulations and hooked them up to an electroencephalogram (EEG) to monitor their mental load while playing. Those using the 2D version did better on knowledge-assessment tests, he found³; the EEG results hinted that students were “overstimulated” by the 3D version. “Students strongly prefer immersive virtual reality, and this leads to significantly higher presence, motivation and perceived learning,” he says of the 3D systems. “But presence does not necessarily always lead to higher learning.”

“So you think, crap, we just did years of work on 3D systems,” laughs Labster’s Bodekaer. But such studies just prove that you have to be careful about how you design each specific learning experience, he argues. “You cannot just take traditional laptop-based sims and port them one-to-one to virtual reality. It’s cognitively more immersive — that has benefits but it means you cannot overload the students.”

Some of Makransky’s other studies showed that virtual-reality sims, whether 2D or 3D, can impart knowledge better than a textbook. “That’s really exciting. That’s exactly the results we were hoping to see,” says Bodekaer.

All this implies that virtual labs can be used to replace some real-world labs, just as Hamadani has done for his students. But replacing the entire lab component for a science degree is unusual. “There are a few other biology degrees offered online, but most are bachelor of arts degrees,” says Amy Pate, instructional designer at

“The real lab is actually very limiting. A lot of it is toxic. We can’t let them blow up things.”

ASU’s School of Life Sciences. From this term, her university is using a set of 30 virtual simulations made by Labster in collaboration with Google Daydream for three of its core lab courses: cell and molecular biology, animal physiology, and ecology — but students need to take a real-world organic-chemistry lab. “The learning objectives and rigour of the course are the same, regardless of the modality in which the student learns,” argues Pate.

There are still fundamental differences between real and virtual labs, of course. The data in simulations tend to be less messy, leading to faster, cleaner learning. And without the distraction of trying to physically wrangle equipment into working properly, students can spend more time understanding the principles behind the science. But this also means they get less practice in working out how and why things go wrong, and have less experience with the arguably useful emotion of frustration. “It’s really different to encountering a conundrum in the lab,” says Linn. “Maybe you put your materials too close to the fridge and

it interfered.” That level of complexity is not available in simulations — not yet anyway.

MIX AND MATCH

For Labster, the future lies in customizable sims that can be adapted to a teacher’s needs, and they are prototyping and testing Lab-Builder with teachers this year. This service lets anyone use Labster’s simulated tools and lab environments to build their own labs, much as gamers use SimCity to build their own cities.

Such flexibility is also emerging as a key driver for edX, which was founded by Harvard University and the Massachusetts Institute of Technology in 2012 to provide access to ‘massive open online courses’. It currently hosts some 2,000 courses with at least 130 institutions. They plan to shake this up, inventing a system that replaces discrete courses with an array of mix-and-match components, from lectures to lab simulations, that can be freely customized to anyone’s needs. “Imagine if you can cherry-pick from 1,600 courses exactly what’s relevant for your students. That’s changing the game,” says Robert Lue, faculty director at HarvardX, who heads the LabXchange project funded by the Amgen Foundation.

Lue’s team is designing new 2D lab sims for their project and hoped to release the first of these as a pilot in September. They aim to focus not on procedures, such as loading a pipette, but on the process of experimental design: choosing an experimental approach, modifying a protocol, analysing the data, working out what went wrong, and doing it all again.

For Lue, these simulations, and the mix-and-match access to pieces of course material, will revolutionize learning. “For the hundreds of millions who’ll never be able to use a \$15,000 PCR machine, they can have the experience of design, failure and redesign,” says Lue. “That’s what’s really getting me out of bed.”

But more than that, he adds, the virtual world lets all students, regardless of their real-world lab access, fail multiple times without cost, try out a limitless variety of variables and procedures, and crank through the whole process of science in a swifter, cheaper and more efficient way than is currently possible.

It is doubtful whether virtual labs will completely replace real bench work in scientific training just yet. But that might not matter. “It’s not an either/or thing. It’s not like virtual labs are going to take over,” says Hamadani. “But when strategically used in the right way, they can improve all kinds of learning outcomes.” And for those who have no access to real labs, a virtual lab is better than no lab at all. ■

Nicola Jones is a freelance reporter based in Pemberton, British Columbia, Canada.

1. de Jong, T., Linn, M. C. & Zacharia, Z. C. *Science* **340**, 305–308 (2013).
2. Bondé, M. T. et al. *Nature Biotechnol.* **32**, 694–697 (2014).
3. Makransky, G., Terkildsen, T. S. & Mayer, R. E. *Learn. Instruct.* <https://doi.org/10.1016/j.learninstruc.2017.12.007> (2017).

PERSPECTIVE



Drawn to science

Teachers do not need training in the arts to create useful drawing experiences for science students, says **Bethann Garramon Merkle**.

People think in images. Indeed, archaeological records show that drawing was the first means of visual representation (see go.nature.com/2qfxe7h). The lines and dots of the earliest rock art — some 64,000 years old — indicate conceptual, creative thinking. Cave engravings in France depict well-proportioned, figurative observations of wildlife made some 26,000–38,000 years ago (below). Today, images remain powerful tools for learning, documenting and facilitating thinking. Taking notes by hand¹, or even doodling, results in better retention of information and higher intellectual engagement with the material than does typing on a laptop. Fundamentally, creativity is a whole-brain process, and artists and scientists use the same parts of their brains to do complex, creative tasks². Ensuring that students understand the value of drawing can help motivate them to draw.



Sketches documenting prehistoric rock carvings from the Grotte de Pair-non-Pair in France indicate the use of perspective in prehistoric art.

When my colleagues try to integrate drawing into their laboratory and field courses, however, they frame their motives more matter-of-factly. For example, one biology-lab coordinator noticed that students mainly interact with specimens by photographing them. She suspected that students did not gain much from taking these photos, on the basis of their exam scores. Her hunch is substantiated by research indicating that analogue note-taking is more effective than using laptops¹, and that students score more highly on exam questions that correspond to topics for which drawing is required. To make students engage more fully with specimens, I worked with the lab coordinator to build drawing into the curriculum. But even without previous training in the arts, teachers can create productive drawing experiences for their students.

Drawing is a scientific tradition that has contemporary validity. Drawing has been an important tool throughout the history of science. Connecting students' coursework to this tradition encourages students to see drawing as a fundamental tool of science. In ancient Greece, for example, protoscientists such as Aristotle drew to help them understand animal anatomy. In Renaissance Italy, Leonardo da Vinci's drawing-based investigations of the erosive force of water predated fluid mechanics by several centuries. In the seventeenth century, Maria Sybilla Merian was the first to study and depict ecologically accurate insect life cycles (above right). Her illustrations definitively countered the prevailing belief in spontaneous generation.

By developing their drawing skills through practice in the classroom, students can contribute to this legacy as well as enrich their own scientific experiences.

Drawing skills can be learnt and taught. Students frequently object when asked to draw because they have little confidence in their abilities and limited training in the arts. Indeed, without adequate basic training, drawing can be a hurdle rather than a tool. However, drawing in science need not be about making great art; the emphasis can be on accuracy, enhanced observational skills and asking deeper questions. Furthermore, the basics of modern drawing are learnt, not inherited. For example, 3D drawing took about 400 years to develop and was not fully understood until the Renaissance. Fundamental skills, techniques and knowledge of different media, such as watercolour, pencil, and pen and ink, can be taught, applied and improved.



Maria Sybilla Merian's seventeenth-century study of the insects of Suriname and their life histories foreshadowed contemporary ecology.

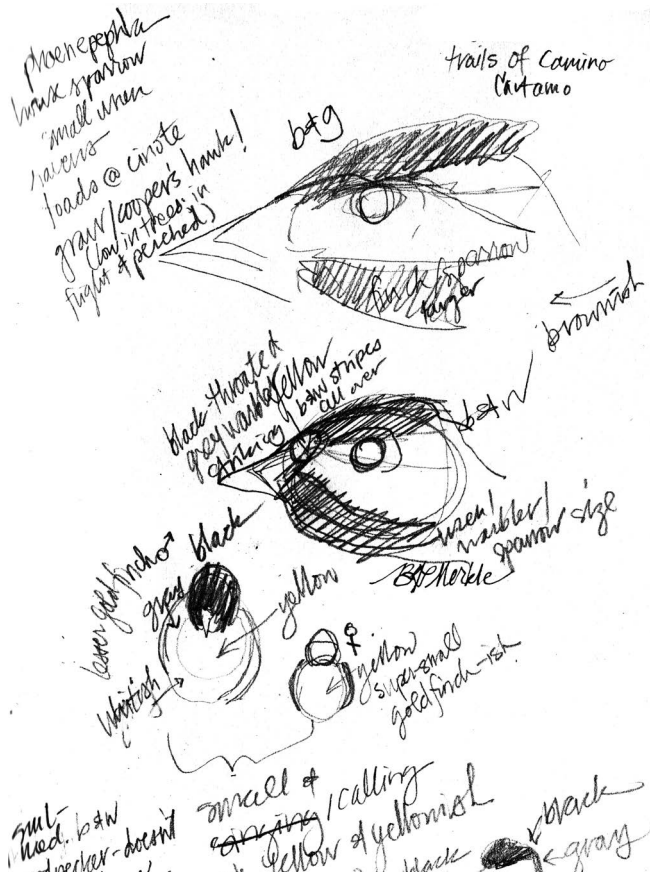
Students' coursework should include training in basic drawing skills and an expectation that students will draw in every class. Teachers who lack the confidence to lead drawing instruction can collaborate with a colleague who has drawing skills or their institution's art faculty. Individuals and companies offer arts-related professional development and classroom-based instruction, including local, professional, science-focused artists. To find such artists in the United States, check the Guild of Natural Science Illustrators' membership list (gnsi.org); other countries have similar professional organizations.

For resistant students, Charles Darwin offers a cautionary tale. Despite encouragement from his colleagues, Darwin refused to learn to draw and always regretted it. In his autobiography, he bemoaned the "irremediable evil" of his "incapacity to draw" and acknowledged that his notes were less useful than they could have been if he had made his own illustrations. Similarly, teachers need to commit to drawing as a demonstration of what they expect from their students. Consider, for example, drawing during lectures, sketching alongside students during dissections to point out features of interest, or even including rough sketches in lecture slides.

The way that drawings are assessed is crucial. As well as serving as a learning tool, drawing can be a powerful mechanism for both formal and informal assessment. In the project Picturing to Learn (www.picturingtolearn.org), drawings by students at Harvard University, the Massachusetts Institute of Technology, Duke University and other institutions were evaluated for the presence or absence of essential aspects of scientific systems and concepts. The drawing assignments and test questions “revealed misconceptions in a way that text does not”, according to one participating professor. For example, asking students to draw a diagram explaining why the sky is blue forces them to demonstrate, beyond memorized equations and keywords, what they do or do not understand.

The type of drawing matters less than the fact that students are drawing. Teachers could ask for a highly polished illustration of a leaf, sketches of a few circles indicating locations at a research site, or a diagram of genetic relationships in a phylogenetic tree. The context will determine how representational or abstract a drawing needs to be. For example, to represent a wolf, a box containing the word ‘wolf’, a stick drawing of a wolf, a caricature, or a more photorealistic picture of a wolf might be appropriate, depending on the assignment³.

Importantly, teachers should evaluate the content and accuracy of such sketches, not their artistry. For example, the sketches below depict birds that I found difficult to identify in the field. Quick and rough, these sketches captured fairly detailed information in the line work and accompanying notes that proved sufficient for looking up and correctly identifying the species later: it was the northern phainopepla (*Phainopepla nitens*). Sketching in this way can enhance learning, observation and drawing skills. For example, these drawings are stronger artistically than the bird sketches I made a decade earlier when I first began to use drawing to understand



Bird sketches by Bethann Garramon Merkle; their detail and accuracy enabled Merkle to identify the species later, when she added the species name and further notes.



Dene First Nation hunters and elders annotated a researcher's drawing to combine traditional ecological knowledge with a genetic study of caribou biodiversity.

ecology. Combining both practice and training can lead to enhanced artistic skill.

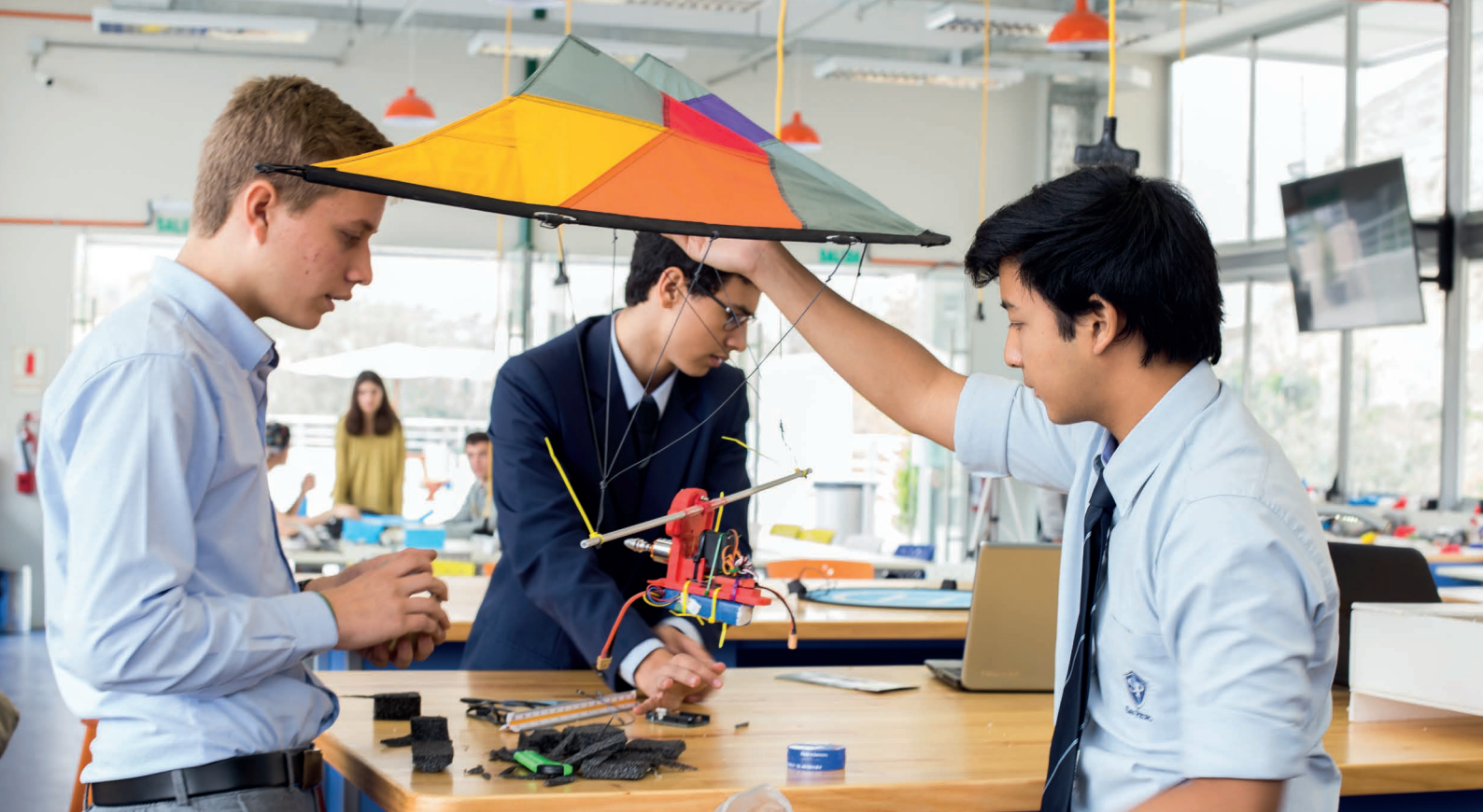
Science teachers should get their students to use drawing as part of informal activities as often as possible. The idea is to gradually build up to low-stakes, graded assessments before ultimately using the students' drawings in high-stakes assessments.

As well as supporting learning outcomes, research indicates that drawing (even without training) can enhance visual-thinking skills, creativity and problem-solving, and can improve science-communication efforts. There is even evidence that collaboration between scientists and artists can result in better science. For example, successful caribou management and conservation in Canada depends on strong relationships between researchers and indigenous First Nations communities. These relationships are often tenuous. Jean Polfus, an artist and ecologist at Trent University in Ontario, used drawings to document Dene First Nation traditional ecological knowledge about caribou phenotypes. Local hunters and elders annotated the drawings (above), which Polfus then revised for them to review. Through this iterative drawing process, elders overcame their reluctance to participate in the research. Polfus was able to incorporate traditional ecological knowledge in genetic analyses of scat samples collected by Dene hunters. Ultimately, these genetic analyses accorded with the phenotypes recognized by tribal elders. Drawings were at the root of this collaborative process⁴.

In similar ways, researchers, instructors and students can use drawing to learn course content and enhance scientific thinking. Sketching, assessing and revising drawings will provide practice in a skill set that has been central to the practice and communication of science for centuries. ■

Bethann Garramon Merkle is an illustrator and researcher at the University of Wyoming in Laramie. She has consulted on the implementation of incorporating drawing as a teaching method in science classrooms throughout North America.
e-mail: bmerkle@uwyo.edu.

- Mueller, P. A. & Oppenheimer, D. M. *Psychol. Sci.* **25**, 1159–1168 (2014).
- Andreasen, N. C. & Ramchandran, K. *Dialogues Clin. Neurosci.* **14**, 49–54 (2012).
- Quillin, K. & Thomas, S. *CBE Life Sci. Educat.* **14**, 1–15 (2015).
- Polfus, J. L. et al. *Ecol. Society* **22**, 4 (2017).



United Technology for Kids sends science students at US universities to Peru to help local schoolchildren to develop their practical skills.

DEVELOPING WORLD

Expanding the reach of science

Science education is helping to strengthen the future prospects of developing countries.

BY NEIL SAVAGE

Guiliana Huerta-Mercado feels that she had a good school education in Lima, Peru. But shortly after arriving at the University of Michigan as a first-year economics student in 2015, she was startled by what she saw. Some of her US friends were already working on projects that involved robots or drones. “I was like, how do you know all these things? You haven’t even started doing engineering,” she says.

The reason they were familiar with robotics, her classmates told her, was that they had gained experience with it at secondary school. Their education had been very different from her own, which was based on book learning but included little practical work. Her school held a science fair each year, which was more than many schools in Peru did, but the only thing she had built was a baking-soda volcano.

So Huerta-Mercado got together with other students at Michigan, the University of California, Berkeley, and the Massachusetts

Institute of Technology (MIT) in Cambridge, to found United Technology for Kids, a non-profit group established to provide students in Peru with exposure to science, technology, engineering and mathematics (STEM). For the past three years, students from the group have travelled to Peru to introduce younger students to these disciplines.

Dozens of organizations around the world are delivering STEM education to people in developing countries, often at the behest of, and with financial support from, those countries’ governments, which see training in science and engineering as a way to bolster the economy. International institutions such as the World Bank and the United Nations Educational, Scientific and Cultural Organization (UNESCO) extol both the economic and the human-rights benefits of teaching STEM subjects to more people. And multinational corporations, eager to have a skilled local workforce where they would like to locate, are supporting such efforts. The challenges of introducing STEM education vary from

country to country, but common problems include a lack of resources, resistance to changing the curriculum, and social inequality, especially for girls and young women.

United Technology for Kids has sent 32 university students from the United States, many of whom have a Latin American background, to spend three weeks teaching schoolchildren in Peru the basics of topics such as electronics, robotics and biotechnology in after-school programmes. When the US students return home, local university students continue to run workshops for the next five months. The children work on projects connected to what they are learning at school. One group, at a school in an arid part of the Andes, built a drip-irrigation system to create a green area where kindergarten-aged children can play. Another used a 3D printer to create an affordable prosthesis for children missing a hand who need replacements as they grow. “We try to get the students thinking about things they can do to solve a problem that Peru has,” Huerta-Mercado says.

So far, the scheme has reached about 1,000 students at 20 schools in Peru, as well as students in Medellín, Colombia. After taking part, 96% of the children said their interest in engineering had increased, and 26% said they had changed their career preference to a job in a STEM field.

TAKING ACTION

The programme fits the approach, generally championed by proponents of STEM education, of ‘active learning’, in which students learn a concept by making use of it, rather than by reading about it in a textbook. It asks them to identify problems and to work out a possible solution. At first, that was a tough sell to students who were accustomed to being told

KATHERINE GOICOCHEA

what they needed to know. “In Peru, students are not used to doing things on their own,” says Huerta-Mercado. They wanted step-by-step instructions on how to do everything.

This is a common problem found in many countries that cling to a textbook-centric, exam-evaluated model of education. “They just wait as empty containers to be filled,” says Alan West, a STEM consultant and former chemistry teacher. West’s company, Exscitec, based in Petersfield, UK, is developing a STEM curriculum for the Ministry of Education and Training in Vietnam, where the problem is entrenched. Teachers in the country, where every school is on the same page of the same textbook on any given day, feel that they must stick to the rules or face losing respect and authority. They are not comfortable with the free-form approach that is common in modern STEM education.

West invited some teachers from Vietnam to visit UK schools to see how they had embedded STEM subjects in their curriculum. Three of the schools did not even issue textbooks, although they were available from the library. “That was quite a revelation for them,” West says.

West is currently training teachers at schools around Hanoi. Local teachers and school administrators felt they lacked the authority to change the way in which students were taught, so the ministry issued orders that let them alter the curriculum. The idea is not only to show teachers how to incorporate STEM learning into their own classrooms, but also to train enough of them so that they can go on to train others.

The emphasis on rote learning in developing countries might seem old-fashioned, but some of it is down to necessity, argues Fanuel Muindi, who founded the US STEM Advocacy Institute in Cambridge, a think tank that promotes STEM education. Hands-on learning is hard to do in a setting in which you lack resources, Muindi says. “Teachers are forced to say, ‘OK, we’re going to learn from the book.’” Access to resources is also a problem in Africa, he adds. Countries such as Kenya, Nigeria and South Africa do well with STEM education, but lower-income nations have a harder time.

In nations with high rates of poverty and poor infrastructure, providing a basic education can be a struggle. It is hard to worry about adding STEM subjects to the curriculum when “people can’t even get to the classroom”, Muindi says.

LOCAL FLAVOUR

Those who hope to expand STEM education in developing countries tend to use ideas and teaching tools from higher-income nations but must adapt them to the needs of the local community. “Whatever STEM education is happening needs to happen in a local context and not copy what’s outside,” says Connie Chow, a biologist who founded The Exploratory in Boston, Massachusetts, an organization that has trained about 700 STEM teachers in Accra. She helps teachers to develop course units that focus on matters of interest in Ghana, such as agriculture and malaria.

Local context can extend to language. Some schools in Vietnam teach in both English and Vietnamese, West says, and he has visited schools in Malaysia and Cambodia at which the teaching is done in English. In Ghana, where the official language is English and tests are given in English, the level of fluency varies and there are hundreds of dialects, Chow says. Most instruction in Peru is in Spanish, says Huerta-Mercado, but some students in rural areas speak the indigenous language Quechua.

Michel DeGraff, a linguist at MIT who directs the MIT-Haiti Initiative, says that students learn best in the language with which they are most comfortable. “How can you expect a child, or any student for that matter, to become fluent in science or math if the language used to teach it is one the student doesn’t understand?” he asks.

DeGraff is from Haiti, where schools teach in French. However, most Haitians, including many teachers, are poor at French, preferring to speak in Haitian Creole (Kreyòl). DeGraff runs a project that introduces Haitian teachers and students to STEM subjects by using instructional tools such as PhETs — game-like computer simulations developed at the University of Colorado Boulder, that are designed to teach concepts from science and maths. When PhETs and other resources are translated into Kreyòl, students take to them more quickly, DeGraff says. “The students get more animated. They ask more questions. They are smarter.”

One hurdle the project faced was finding words in Kreyòl for certain scientific concepts. DeGraff and his colleagues often adapted terms from French or English, but sometimes they repurposed existing words. To translate the word ‘torque’, for instance, they used the Kreyòl word *tòday*, which is the motion of wringing out a wet cloth.

In 2015, the government of Haiti announced a policy to educate students using Kreyòl, but DeGraff says it often fails to follow through on such promises. In other countries, teaching tools are available in a variety of languages. Some PhET modules have been translated into languages such as Afrikaans and Welsh.

GENDER EQUITY

A major focus for people helping to improve STEM education in developing countries is making sure that it reaches girls as well as boys. A 2017 report¹ by UNESCO found that, worldwide, only 35% of students enrolled in STEM courses in higher education are female, and only 28% of researchers are women.

“For sustainable development, we need more scientists and more woman scientists,” says Alessandro Bello, a social scientist who leads UNESCO’s STEM and Gender Advancement project in Paris. “All the important jobs are going to be related to STEM.” Having enough people to fill those jobs will require women to be educated in STEM subjects. It is also a matter of human rights, he suggests: being educated in such subjects gives women

access to the income and status that flows from jobs in STEM fields.

Addressing the problem is difficult because there are few data that show which interventions are effective, says Ana Maria Muñoz-Boudet, a social scientist at the World Bank’s Poverty and Equity Global Practice. She and her colleagues looked at 2,000 papers published between 2000 and 2016 on STEM education for girls. Only about 250 discussed policies or interventions, and only 19 included rigorous reports on the outcomes of interventions². Most were from the United States or Europe, not developing countries. “We don’t know what is working and what is not,” she says.

But providing girls with female role models is likely to encourage them into science careers. A 2013 survey³ by researchers at Florida Gulf Coast University in Fort Myers found that girls who attended STEM workshops led by women reported an increased interest in those fields. And a 2018 survey⁴ by Microsoft of more than 11,000 girls and women in 12 European countries found that 41% of girls who had such role models, whether real or fictional, were interested in STEM subjects, compared with only 26% of girls who lacked such role models.

The emphasis on STEM education by governments in developing countries comes from the idea that creating a workforce with technical skills will make such countries more attractive to multinational

“The students get more animated. They ask more questions. They are smarter.”

corporations looking for places to locate offices and factories. Companies from Japan, South Korea and Singapore are now investing heavily in Vietnam. Chow’s

group receives funding not only from Ghana-based companies, but also from the High Commission of Australia because Australia is involved in mining in West Africa. Multinational corporations “want to use STEM education as a way of growing local talent”, says West.

The aim of STEM education should be wider than encouraging more students to become scientists, Muindi suggests. It teaches students to think more systematically, he says, which helps them to analyse problems and find solutions. This skill is equally applicable to careers in business or politics. “My goal is not for people to become scientists,” he says. “I see understanding science as a gateway to success.” ■

Neil Savage is a science writer in Lowell, Massachusetts.

1. UNESCO. *Cracking the Code: Girls and Women’s Education in Science, Technology, Engineering and Mathematics (STEM)* (UNESCO, 2017).
2. Henninger, N. J., Muñoz-Boudet, A. M. & Rodriguez-Chamussy, L. *Evidence From Policy Interventions Promoting Girls’ Participation in STEM: A Systematic Review* (in the press).
3. Dubetz, T. & Wilson, J. A. J. *STEM Educat.* **14**, 41–47 (2013).
4. Microsoft & KRC Research. *How Role Models Are Changing the Face of STEM in Europe* (Microsoft, 2018).

Drive for diversity

Physics struggles to attract and retain graduates from underrepresented groups, but changes to the way it is taught may help to close the gap.

BY NICOLA JONES

At Reed College in Portland, Oregon, a young woman of colour — a first-year student interested in physics — made an appointment to see Mary James, the college's dean for institutional diversity. "She said: 'I heard there's an African American physicist on campus and I just wanted to meet you'," says James. In their conversation, the student recalled standing at the blackboard the previous week to work on a problem and suddenly realizing that all her group partners were white men. She thought that if she failed they would think that women of colour couldn't hack it. "She said: 'I knew I shouldn't be thinking that. But I couldn't help it.' It was a classic stereotype threat," says James, who handed the student a book from her bookshelf on the phenomenon. "This is a really powerful thing, and it really impedes performance."

James, who now chairs an American Institute of Physics (AIP) diversity task force, is a member of a small club. She was one of only 66 black women who earned physics doctorates at US universities from 1973 to 2012, compared with more than 22,000 white men and over 2,400 white women.

Diversity is an issue across all the sciences, but in the United States, physics (along with maths and engineering) is near the bottom of the pile. The National Science Foundation reported¹ that in 2014, people from underrepresented groups (black people, Hispanic people and American Indians or Alaska Native groups) earned about 20% of science and engineering bachelor's degrees awarded in the United States. The AIP also drilled into the numbers and found² that although underrepresented-minority (URM) people make up about one-third of the US population, they are awarded just 11% of physics bachelor's degrees and 7% of PhDs (that's a meagre 60–70 students per year). The percentage of faculty members who are African American actually decreased slightly from 2008 to 2012. Although women obtain just over half of the science bachelor's degrees and PhDs in the United States, they get only 20% of the physics degrees (down from a peak of 23% in 2004).

The reasons for these disparities are many and varied, and leaks happen at every stage of the educational pipeline from elementary school upwards. URM populations are more likely to be economically disadvantaged, for example, leading to poor access to good education and a resulting lack of opportunities.

The lack of role models is also a huge issue, says Jami Valentine, a patent examiner at the US Patent and Trademark Office and an advocate for African American women in physics. "Too many professors have never taught an African American in a graduate course, and they likely never had a colleague who was African American," says Valentine, a former board member of the National Society of Black Physicists. "Students need to see physicists who look like them."

Physicist Mary James chairs an American Institute of Physics diversity task force.

James' task force at the AIP convened in December 2017 and will focus specifically on African Americans in undergraduate physics. The choice of target is strategic, as it highlights an egregious lack of progress: a smaller proportion of physics and astronomy bachelor's degrees are awarded to African Americans today than two decades ago. Over the next two years, the task force will survey students and visit schools to find best practices and develop recommendations to increase the representation of African Americans in these disciplines.

But everyone acknowledges that work is needed across the board to close the gaps for all URM students at all levels of education. "We need to focus on all areas," says Ximena Cid, a physicist at California State University in Dominguez Hills, who publishes on diversity issues in physics. "At each level we're losing people."

HIGH INCLUSIVITY

One wall of the Manor New Tech High School in Texas features a mural painted by the students. In cool blues and greys, it depicts the unformed shapes of a boy and a girl entering a chemical apparatus and bubbling out of the pipeline at the other end. When the school opened in 2007 on the outskirts of Austin, principal Steve Zipkes told researchers at George Washington University that "it was just to get our kids to go to college". Of the district's students — 79% of whom are economically disadvantaged and 24% are African American — only 40% were completing high school and just 15% went to college. By 2010, Manor New Tech was sending more than half of its students to four-year postsecondary institutions compared with a national average of 28%.

Sharon Lynch, a science-education researcher at George Washington University, holds up Manor New Tech as an exemplar of an inclusive science, technology, engineering and mathematics (STEM) school. Lynch and her team are starting to show that such schools are having great success in ensuring that URM students get the education they need to set them up for science at university.

In the United States, says Lynch, URM populations typically face a double-whammy. "Poor kids living in poor neighbourhoods are really underserved by school funding," she says. For example, almost one in five African American high-school students attends a school that does not offer any advanced placement courses. But even in schools that do, says Lynch, URM students are more likely to be placed on courses that are less academically demanding than are white, middle-class children.

"For minority kids, access isn't enough," says Lynch. If a poor black girl with an interest in STEM attends a good school, theoretically she has access to top-stream classes, but in reality, says Lynch, that access is difficult to obtain. "You could go into any high school in my area and tell with exact accuracy the level of a course from the proportion of brown and

black students in a classroom," she says. "The only high schools that I have seen completely dismantle this practice are the inclusive STEM high schools."

Unlike science-focused schools that aim to attract high performers, inclusive STEM schools admit students on the basis of interest, rather than test scores. They are public schools, often with no special entrance criteria, and some even use a lottery system. Their mandate is to give all of their students college preparatory work, rather than just some of them. "These schools don't just provide access, they make sure you're having that experience," says Lynch.

The schools emphasize many of the things advocated by education researchers for all groups, all ages and all subjects of study. There is a lot of problem-based learning, for example, and a strong sense of community, and they incorporate new technologies into everyday activities. Lynch's team sat in on a chemistry class at Manor New Tech in which students were designing a gas canister for use in a biodome on the Moon. Less than 15% of class time was spent on instruction from the teacher, and some of that was specifically requested by the students.

In 2012, Lynch and her colleagues started studying eight high-performing inclusive STEM schools to understand why they are so successful. Her co-investigator Barbara Means from SRI International, a non-profit research organization based in Menlo Park, California, has crunched the numbers for about 50 inclusive STEM schools in North Carolina and Texas.

The study found that these schools are clearly living up to their 'inclusive' mandate. For example, half of the high-school graduates in the North Carolina schools were African American, compared with just 9% in the 2013 class of the selective North Carolina School of Science and Mathematics. In both North Carolina and Texas, most of the students came from low-income homes — a proportion that exceeds, or is on a par with, the averages in these states.

Students who went to these inclusive STEM schools graduated with stronger attitudes about, and more career interest in, science than average state students. Lynch cites a longitudinal study that tracked the STEM schools' students two years after graduation: "African Americans, Latinos and girls all do better on many measures than their counterparts in comprehensive schools," she says. "More kids are going to college and more kids are staying in college."

Even at schools that do not focus specifically

on STEM, physics teachers have access to a host of approaches to make them more inclusive, many of which boost test scores and attitudes among URM students in particular.

DON'T GIVE UP

Eugenia Etkina grew up in Moscow, where she trained as a physics teacher. She moved to the United States in 1995 and now researches physics education at Rutgers Graduate School of Education in New Brunswick, New Jersey. In the 1990s, she saw that most high-school and undergraduate physics classes were taught using the 'predict, observe, explain' model. This might sound like a sensible way to learn physics, but Etkina argues that it damages women and URM students.

Much of the time, Etkina explains, physics predictions based on intuition are wrong. If a teacher asks what falls fastest, something heavy or something light, for example, most people would say something 'heavy', based on everyday experiences with rocks versus leaves, say. When the teacher whips out two equal-sized, differently weighted balls and proves you wrong, it seems like a trick. The intended effect is to surprise the student, making them curious and keen to solve a mystery. Why do these balls fall at the same speed? Is it really exactly the same speed? And at what point does air friction make a difference? But instead it often makes students feel threatened and defensive. "Then they think: 'I'm stupid — I don't belong here,'" Etkina says.

Groups that are already predisposed to thinking that they don't fit in, because of their minority status or various cultural factors, are hit hardest by such blows. URM people have been shown to face stigma and stereotypes that can affect their self-confidence³ and their performance in science classes, for example, and women tend to judge themselves more harshly than men judge themselves⁴. "As a woman, I realized what I was putting my students through," Etkina says.

Etkina dedicated herself to designing some alternative teaching philosophies to get around this problem specifically for physics, and developed the Investigative Science Learning Environment (ISLE). In this system, she explains, students are given the chance to observe a phenomenon before coming up with testable observations. "We call them 'crazy ideas' so it's fun to rule them out," she adds. Maybe, for example, the speed of an object's fall depends on whether it is made of rock or rubber. This promotes a 'mistake-rich' environment in which the fear of failure is reduced.

Etkina says that more than 1,000 teachers have used the ISLE approach in their high-school and undergraduate physics classes, using a textbook and materials designed to support the approach. Suzanne White Brahmia, a physics-education researcher at the University of Washington, integrated ISLE into her first-year physics course⁵ at Rutgers

"Too many professors have never taught an African American in a graduate course, and likely never had a colleague who was African American."



Ximena Cid (second left) studies diversity issues at California State University in Dominguez Hills.

University in 2001 and credits it in part for a huge change in the demographics of her students. The percentage of URM students who went on to complete a STEM degree in under six years jumped from 8% in the late 1980s to 58% in 2008.

Geraldine Cochran, who studies physics education and teaches at Rutgers, started using ISLE in 2017. It has many good qualities, she says, including “creating a culture where it’s okay to make mistakes, highlighting strengths instead of weaknesses.”

The ISLE approach is just one of many teaching innovations in STEM that seem to have a disproportionately positive impact for women and URM students. Another of these, the Student Centered Active Learning Environment with Upside-down Pedagogies (SCALE-UP), was developed in the physics department at North Carolina State University (NCSU) in Raleigh. This approach involves getting students to watch lectures or read books before coming to the lesson so class time can be spent on discussion and problem solving. The teachers who use SCALE-UP report that their students become more comfortable with the idea of not understanding something at first glance. They tend to feel less isolated and are less likely to think they are alone in not ‘getting it’. This in turn reduces failure rates, particularly for URM students. In a study of more than 16,000 students taking introductory calculus-based physics courses at NCSU over a five-year period, students in a more-conventional lecture-style class were about 2.8 times more likely to fail in tests than SCALE-UP students. That proportion shifted to 4.7 for women and 3.5 for African Americans.

But assessing the effectiveness of such programmes is difficult because they are confounded by other variables, including societal evolution. There are established ways of

assessing whether changes to undergraduate education improve student understanding, including simple things such as class size or course content, but even these are subject to bias. According to Cid, 63% of college students represented in the US physics-education research literature are white, compared with 45% for all college-bound students. This imbalance stems from the fact that most of the studies are done at selective, top-tier research universities that have fewer URM students. “Effectively, the physics-education research community has inadvertently cherry-picked its data,” wrote Cid in a 2017 paper⁶. They have created an accidental focus on “well-prepared calculus-based students with relatively homogenous and privileged backgrounds.”

TOP TIER

If students can be attracted to university physics, and stay long enough to earn a degree, the final educational hurdle is to get them into graduate work. The American Physical Society’s Bridge programme aims to do just that, working with more than 35 US institutions to improve admission and retention rates for URM students in graduate physics programmes.

One of the main effects of the Bridge programme is to counteract biases introduced by the Graduate Record Examinations (GRE), a standardized test required for admission into most graduate schools in the United States. A survey of about 150 US physics graduate programmes (out of about 200 in the United States) showed that more than one-third use GRE scores to make admissions decisions. But research has shown that GRE scores do not relate to success in completing a PhD. They do, however, correlate strongly with race and gender⁷.

The result of using GRE scores as cut-offs, argues Casey Miller, associate dean for research and faculty affairs at the Rochester

Institute of Technology, is “a glass ceiling erected by the lopsided treatment of minorities and women before they even set foot in grad school.”

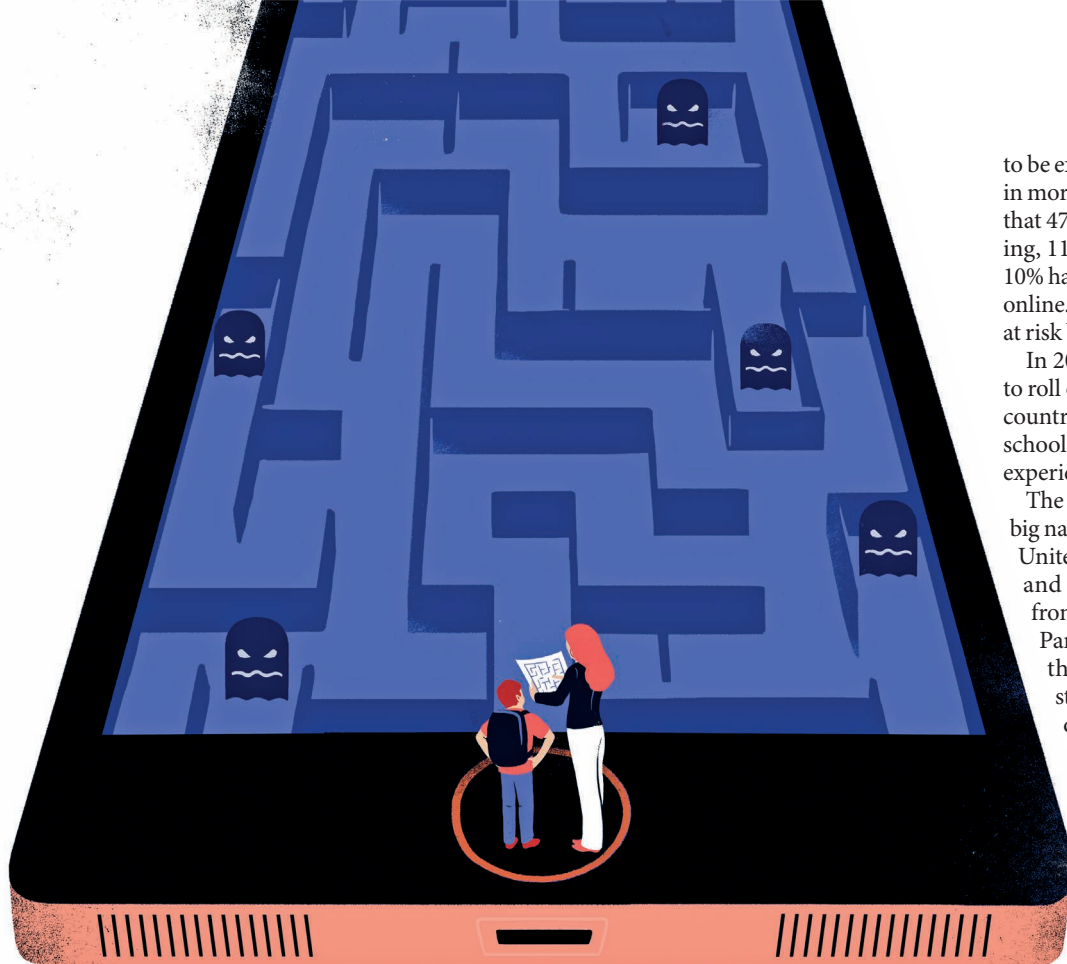
Bridge programmes encourage universities to consider other qualities, such as a candidate’s devotion to the subject, motivation to work hard, creativity, tenacity and drive. They also provide mentoring for students once they are in graduate school, along with some financial support. Since its launch in 2012, the APS Bridge programme has helped 129 students enrol for PhDs. If all these students complete their studies, this will double the percentage of PhDs awarded to URM students in the United States. Individual universities have seen similar gains. Ohio State University, for example, made a conscious effort to increase diversity in its physics PhD programme, in part through a Bridge programme. It saw URM students rise from less than 5% of its domestic students in 2012 to almost 20% in 2017.

The effects of such programmes might also be self-reinforcing. Just as a lack of role models tends to undermine the retention of URM students in physics, their presence can have disproportionately large effects. Making role models more available and more visible is a major part of the work of the National Society of Black Physicists, says Valentine. The organization hosts an annual conference, partly to help reinforce the sense that it is normal for African Americans to do physics, and also to forge connections for mentorships. Its upcoming conference, taking place this November in Columbus, Ohio, is expected to bring together more than 500 African American students and professionals.

Many programmes are aimed at targeting funding and efforts specifically towards URM populations, but others are simply about changing the way physics is taught to make it more inclusive from the outset — not only to include all races and genders, but also all learning styles and a more diverse array of talents. Students shouldn’t be thinking ‘I don’t belong here’, says James, “even if they’re likely to overhear that in the cafeteria”. Programmes that lessen the fear of speaking up, promote teamwork and self-satisfaction, and chip away at stereotypes, are bound to help the whole of science. ■

Nicola Jones is a freelance journalist based in Pemberton, British Columbia, Canada.

1. *Women, Minorities, and Persons with Disabilities in Science and Engineering* (National Science Foundation, 2017).
2. Ivie, R. *Beyond Representation: Data to Improve the Situation of Women and Minorities in Physics and Astronomy* (American Institute of Physics, 2018).
3. Correll, S. J. *Am. Sociol. Rev.* **69**, 93–113 (2004).
4. Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L. & Espinosa, L. L. *Res. High. Educat.* **50**, 189–214 (2009).
5. Brahmia, S. W. *AIP Conf. Proc.* **1064**, 7 (2008).
6. Kanim, S. & Cid, X. C. Preprint at <https://arxiv.org/pdf/1710.02598.pdf> (2017).
7. Miller, C. W. *APS News* **22** (2), The Back Page (2013).



SEBASTIEN THIBAUT

DIGITAL EDUCATION

The need for digital intelligence

The 'digital intelligence quotient' aims to prepare children for the dangers of the online world.

BY DALMEET SINGH CHAWLA

Children between the ages of 5 and 16 typically spend more than six hours a day in front of a screen. That rises to around eight hours for teenage boys, according to a 2015 report¹ from the UK-based market-research agency Childwise. But spending so long online may impair their ability to recognize emotions² and could expose them to cyberbullying and sexual exploitation.

"I find that quite scary," says Gemma Derrick, a higher-education researcher at Lancaster University, UK. The American Academy of Pediatrics echoes this concern and recommends³ that children spend less than two hours a day with screens.

But instead of minimizing children's exposure to the dangers they may face online, one initiative is trying to build up their digital resilience, teaching children how to deal with those

challenges. The 'digital intelligence quotient' (DQ) helps to provide 8- to 12-year-olds with the skills they need to thrive in the digital economy. Some schools already teach children to use digital technology by integrating computers and tablets into the classroom and teaching them how to code. They should also teach them how to use it safely, say DQ's supporters.

In 2015, Yuhyun Park, who devised the concept of the DQ, founded the DQ Institute, a think tank that aims to prepare children to be safer online. It is often assumed that technology has closed the digital divide and addressed social inequality, but Park's research suggests otherwise. A report⁴ released by the DQ Institute in 2018 studied nearly 38,000 children aged 8–12 in 29 countries and found that more than half are exposed to cyber risks such as cyberbullying, video-game addiction, sexual grooming and sexual behaviour. Children in developing countries are 1.3 times more likely

to be exposed to cyber risks online than those in more tech-savvy nations. The report found that 47% of children experienced cyberbullying, 11% were addicted to video games, and 10% had offline meetings with people they met online. It says that 390 million children will be at risk by 2020.

In 2017, the DQ Institute launched a plan to roll out DQ programmes to more than 100 countries by 2020. The primary target will be schools and education ministries in countries experiencing rapid digital transformations.

The DQ project has joined forces with some big names, including Google, Twitter and the United Nations children's charity UNICEF, and it has received government funding from Singapore and Mexico. The goal, says Park, is to help governments understand the level of digital citizenship among students and teachers, and to help them develop their own DQ curriculum within three years.

In South Korea, UNICEF is supporting the introduction of DQ in schools to help children manage their screen time and boost their critical thinking and technology education. "Korean children spent too much time on screens," says a representative for UNICEF Korea, and that results in a lack of both sleep and physical activity. UNICEF Korea says that DQ skills are associated with critical thinking, which is essential for science education.

ONLINE SAFETY

The DQ has similarities to the intelligence quotient (IQ), emotional quotient (EQ) and social quotient (SQ, the ability to respond effectively after reading a person's behavioural cues and emotions). All of these are needed to navigate the digital world, says Park.

A child's DQ score is calculated from categories that are considered essential to a safe and healthy digital life. The test uses a game-like atmosphere to evaluate how well children understand the importance of their digital identity, privacy management, their online footprint, critical thinking, digital empathy, cybersecurity, cyberbullying management and screen-time management. Much of the information online is unfiltered and uncensored, so DQ programmes aim to provide children with an internal filter, says Park. They teach children how to evaluate online information, avoid dodgy websites, check multiple sources and identify trustworthy sites.

As with IQ, the average DQ is 100 with a standard deviation of 15. A DQ score above 115 is excellent, and children below 85 are considered at risk. A high DQ score means that children are well prepared to use technology responsibly. Park says that a child with the average DQ score of 100 has a 56% chance of being exposed to cyber risks, but increasing the score to 110 decreases the risk to 40%. Raising it to 120 cuts the risk to 28%, she



Yuhyun Park founded the DQ Institute to help children learn the skills they need to stay safe online.

says, and raising it to 130 cuts it further to 18%.

Park says that digital intelligence consists of three components: online citizenship, creativity and entrepreneurship. Online citizenship gives a student the basics needed to operate in the online world ethically, safely and responsibly. Creativity enables them to turn ideas into reality by using skills such as coding and robotics, and entrepreneurship helps them turn a creation into something with economic or societal value. So far, however, the DQ team has developed a test for only the citizenship level, says Park.

INTO EDUCATION

Although DQ tests can be completed at home, Park thinks that teachers can help by monitoring the child's progress, answering questions and consolidating learning after completion of the programme. In a 2016 pilot study, some Singapore schools incorporated DQ into their curriculum. Some children completed the programme on school computers with teacher supervision, whereas other schools let children do it at home without input from teachers. Students at schools that were more involved achieved higher DQ scores.

Samson Wong, who teaches information technology at Man Kiu Association Primary School in Hong Kong, says that the DQ programme in his school "provides comprehensive information and useful skills to instruct our children how to use the Internet wisely and safely".

The number of jobs that rely on technology will increase, so schools have a responsibility to teach students how to use it safely and responsibly, says Sandeep Atre, founder of Socialintelligence, a company based in Indore, India, that runs online courses and workshops on social and emotional intelligence. "DQ will be the most logical choice for us as society to roll out in schools and colleges," he says.

Some schools in Singapore already teach DQ as part of the cyber-wellness curriculum, and in Australia it falls under the technology syllabus. Other schools use it in their ethics

and character-development classes, says Park.

Maimoonah Abdul Malik, cyber-wellness coordinator at Endeavour Primary School in Singapore, says her school uses DQ because it is a good way to teach digital citizenship. But she would like to see "different age groups of students cover different skill sets and attain mastery of all skill sets by the time they graduate from primary school".

Alexander Ray Johnson, technology coordinator at the American School of Bombay in Mumbai, would like a DQ programme tailored for high-school pupils. "All those same skills need to be taught and re-taught and reviewed as the students get older," he says.

Some aspects of digital intelligence may be useful for other purposes. Critical thinking, for instance, is essential for those studying science, technology, engineering and mathematics (STEM). "Scientific thinking and critical thinking are inseparable," Park says. Derrick agrees: "If they can develop this skill early, it will be easier for them to consider a STEM career."

Critical thinking will help children succeed in STEM, says Sonia Livingstone, a social psychologist at the London School of Economics. "I am hopeful that kids' enthusiasm for technology — especially in its more creative, expressive and participatory forms — could be harnessed by schools and non-formal learning sites, including online, in ways that will transfer to STEM learning more widely," she says.

In 2017, Livingstone and colleagues wrote a report⁵ for the UK government exploring the opportunities and risks faced by children online. "The report showed that the risks generally affect only a minority of children, but so, unfortunately, do many of the opportunities," Livingstone says. "If the online world is really to stimulate STEM learning, especially going beyond rote learning and the provision of one-way information, then children must be much freer to explore, experiment and engage with the online world."

Some people are concerned, however, that parents and teachers will become complacent

if a child has a high DQ and treat it as a substitute for parental control. "If they say I don't need to restrict my child's screen time any more because they've got resilience, that's a bit worrying," Derrick says.

She thinks it will be more effective to restrict children's access to social media, which seems to be the gateway to problems online. She adds that social-media providers should take more responsibility for protecting the information of younger people who are more at risk. Some schools and colleges have banned computers in the classroom, but Atre thinks that banning technology in schools is only a temporary fix.

A TIME OF CHANGE

Park agrees that limiting children's access to the Internet and social media in such a hyper-connected world will be difficult. Instead, she says, we need to equip children with the skills they need to be resilient in the digital world, and that's where DQ programmes can help.

How effective DQ is depends on whether it is used appropriately, says Jason Nurse, a lecturer in cybersecurity at the University of Kent, UK. "I think DQ has great potential for change and digital empowerment in children as well as adults," he says.

But Whitney DeCamp, a sociologist at Western Michigan University in Kalamazoo who has studied how violent video games affect behaviour, doubts that the problem DQ is trying to solve is as big as Park suggests. He is concerned by the "alarmist language" of the DQ Institute's report. He also thinks the DQ categories are too broad, so less-risky behaviours are lumped into the same category as more-harmful ones. For instance, he says, a child visiting a site with sexual content may be placed in the same category as one talking to strangers online about sexuality, even though the latter is far more risky. DeCamp says it may be more useful to look at the DQ categories separately, rather than as a combined score.

The digital world is evolving quickly, points out John Mayer, who studies emotional intelligence at the University of New Hampshire in Durham. While we introduce measures such as the DQ, artificial intelligence and augmented and virtual reality are creating environments that humans have never experienced before. Including DQ education in schools could help children navigate this complicated new world, but education needs to keep up with the pace of change in society, he says: "There are very few pieces of this puzzle that are holding still." ■

Dalmeet Singh Chawla is a freelance science journalist based in London.

1. *Connected Kids: How the Internet Affects Children's Lives Now and Into the Future* (Childwise, 2015).
2. Uhls, Y. T. et al. *Comput. Hum. Behav.* **39**, 387–392 (2014).
3. American Academy of Pediatrics *Pediatrics* **132**, 958–961 (2013).
4. *Outsmart the Cyber Pandemic* (DQ Institute, 2018).
5. Livingstone, S., Davidson, J., Bryce, J. & Batool, S. *Children's Online Activities, Risks and Safety* (London School of Economics, 2017).



Beating the odds to secure a permanent contract

Six early-career researchers offer advice on how to secure a permanent contract in academia, and then make the most of it.

Researchers hoping to start their own labs face long odds. A dramatic rise in the number of PhDs awarded each year has significantly reduced the chances of landing a permanent academic contract (see 'Long odds'). Data are sparse, but in the United States, for example, the number of available faculty biomedical positions has fallen since 1980, whereas the number of people who have graduated with a PhD has increased by 60% (N. Ghaffarzadegan *et al. Syst. Res. Behav. Sci.* **23**, 402–405; 2015).

The resulting competition for jobs breeds intense pressure that can take its toll on mental health. A paper published in *Nature Biotechnology* earlier this year reported that graduate students are six times more likely than the general population to experience

depression or anxiety (T. M. Evans *et al. Nature Biotechnol.* **36**, 282–284; 2018). And *Nature's* 2017 graduate survey found that 12% (of 5,700 respondents) had sought help for anxiety or depression caused by their PhD studies (see *Nature* **550**, 549–552; 2017).

The usual advice for people seeking permanent academic positions is to get a good mentor, build up a solid network, publish plenty of papers and hope for a healthy dose of good luck. That advice, once intended to help potential new faculty members gain an advantage, are now minimum standards. A tighter job market demands some updated approaches.

Nature asked six young faculty members for their advice on how to prepare for a permanent position in academia, and how to make the most of one when it comes along.

VICTORIA RUIZ

Design and follow a strategy

Biologist at St Francis College, New York City.

Consider the type of academic you want to become: would you prefer to work in a research-focused institution, where most of your income is funded through external grants, or a teaching-focused institution, in which your salary is paid only if you teach alongside your research? Pick a postdoc position that aligns with that professional goal. ►

► If you are interested in teaching, you will need to demonstrate knowledge in pedagogy. Acquire a teaching certificate, a teaching assistantship or a part-time adjunct position on a fixed-term contract. If you already know you are interested in both teaching and performing research, you can apply for career-development awards that provide mentored postdoctoral research experience alongside teacher training.

Before beginning your job search, prepare statements about your research and your teaching. The research one should include details of your accomplishments and current work, as well as what you would like to achieve in the future. Distinguish your work from your mentor's because it will be difficult to obtain grants if your proposals are too similar to theirs. The teaching statement should discuss your academic mission and educational values, as well as provide an overall narrative of your tutoring style and process.

The search for an academic position is overwhelming, and can be discouraging. For me, however, that career path is one of the most satisfying. Following a strategic plan can help to make it happen.

KIRAN RAOSAHEB PATIL

Work out your numbers

Systems biologist at the European Molecular Biology Laboratory, Heidelberg, Germany.

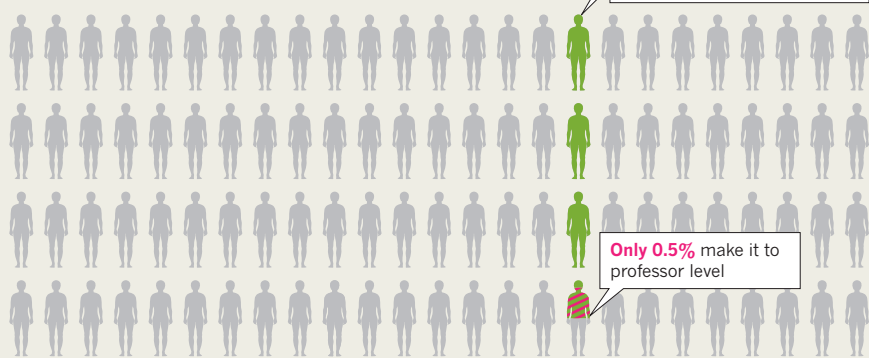
When struggling to find funding for my new lab in 2007, I decided to make life simpler for evaluators trying to compare my application with scores of others. I summarized my research in simple numbers: putting it in terms of publications, citations and the number of times someone requested software I had made.

Don't forget to include the less-obvious achievements. For example, when I was applying for my first faculty position, I had just finished my PhD and I didn't have too many publications. It was unlikely that I would stand out from my publication metrics alone, so I highlighted the number of different research areas in which my PhD work was being used.

Numbers can also act as motivators. Contrast "we aim to quantify a large number of metabolites" with "we aim to quantify 1,000 metabolites". The 1,000 not only gives a better idea of the project to the evaluator but also provides motivation — a sense of challenge and adventure — during the actual execution. And this motivation momentum is a fabulous thing to pass on to your team.

LONG ODDS

Only a small fraction of PhD graduates secure and remain in a permanent research position at a university. Even fewer become full professors, according to a 2010 report from London's Royal Society.



Not everything can be counted, of course, and there is more to research than numbers. But any fractional advantage can make a difference in the current competitive landscape. May the numbers be with you.

AGNIESZKA WYKOWSKA

Ride the wave of uncertainty

Principal investigator and leader of the Social Cognition in Human–Robot Interaction laboratory, Italian Institute of Technology, Genova.

Academic life is a bit like surfing. A surfer's lifestyle provides excitement in pursuit of a never-ending summer and the perfect wave. A scientist's career can also be fuelled by fun, as we explore curiosity-driven questions and travel from conference to conference.

The parallels continue when one looks beyond the positive. Both surfing and academia require extremely hard work, a lifetime of training and commitment, and a huge dose of passion. Succeeding in both requires talent and dedication. But there is one more crucial skill that characterizes both academics and surfers: the ability to ride an uncertain wave. For scientists, the uncertainty lies in day-to-day practice as much as long-term planning.

In the day-to-day, scientists must learn to cope with the unpredictability of experimental results — whether they come out as expected or not. This might affect scientific output, such as publications. This, in turn, affects evaluations and grant proposals that determine the next job contract.

Those short-term events can make the future uncertain. A young academic needs to accept leaving comfort zones over and over

again. They are like nomads — moving from country to country to continue with their science, uncertain of the form and frequency of communication with friends and family.

It can certainly be fun to move between countries and experience new job environments. It is a great experience to meet new people, work in different labs, learn about new cultures. But as time goes by, we also long for some stability and being able to plan more than two years ahead.

Rather than fighting the academic uncertainty, young scientists should try to embrace it. Stability is on the horizon as one becomes more senior. But to reach that goal, as a young academic, it is better to be ready to surf the scientific waves.

MUIREANN IRISH

Make peace with rejection

Cognitive neuroscientist at the University of Sydney, Australia.

Because preparing a publication represents years of hard work, rejection often feels like a personal attack. In my field, journal acceptance rates hover around 20%, and success rates for the two major national government funding bodies are no better. Statistically speaking, rejection is the norm.

I've developed methods to process rejection and learn from it. First, I give myself time. Some scientists — myself included — will need to read the letter, get angry and then complain privately and bitterly about the reviewers until they feel better. After that, I do nothing for at least a week — I simply try to let the dust settle and wait to review the comments when things are calmer.

Rejection is not personal. Perhaps we

misjudged the suitability of a paper for a particular journal, over-interpreted the novelty of our findings, or attempted to publish prematurely. There are similar reasons for rejections of funding applications.

Once the emotional reaction has subsided, discuss the review with your peers. In my lab group, we share our peer-review experiences, which helps to normalize the rejection. By openly sharing that my papers have been, and will continue to be, rejected, I hope to send a clear message to my students that rejection is part and parcel of academia and the world does not end when a paper is rejected.

Finally, never allow your self-worth to be determined by metrics. Academics by nature ascribe to high standards, and to be informed that your work is not good enough can feel like a personal failure. So many factors influence decisions on papers and funding applications, including timing, journal space, funding priorities and, sometimes, just pure luck.

Rejection is the norm, but it is not the end.

KEIVAN STASSUN

Build more than good science

Physicist at Vanderbilt University, Nashville, Tennessee.

Like all people, I think that scientists are at their best — both in and out of the lab — when they feel that they are living lives of passion, purpose and meaning.

Of course, for many of us, the science itself is a passion. But there is also something to be said for having a mission to accomplish something bigger than publication, tenure or even research.

In my own life and career, I have made diversity and inclusion my extra calling. As a first-generation Mexican American from a very-low-income background, working to open doors for others who are underrepresented in science is deeply fulfilling. Even so, there are times when doing so goes against the advice of colleagues, who worry that I am doing that work in place of my science.

But I am convinced that having both focuses has made me a better, more fulfilled person as well as a better, more productive scientist.

While I built my lab I also created links between Vanderbilt and nearby Fisk University, a historically black university, through which more students from underrepresented groups could gain access to PhD courses. In a sense, I then had two opportunities — in my research and outside of it — to experience progress, and this doubled

the reasons for being excited to go to work every day.

Passion and persistence together can lead to the greatest reward of all. Soon after I was awarded tenure, a mentor told me that while I had up to then been focused on building my reputation, now it was time to begin constructing my legacy. That advice has stayed with me. I can see even more clearly now how my two missions working together can make purpose and meaning — and ultimately legacy — possible.

GOSIA TRYNKA

Allow yourself time

Leader of immune-genomics group at the Wellcome Sanger Institute, Cambridge, UK.

Last August, as I planned a seminar for PhD students, I panicked when I realized how little I needed to update my slides. My research had not progressed since the previous year's session. The next day, I shared my dismay with my PhD supervisor. "That's OK," she said, "give yourself five years to evaluate your performance."

She was right. It takes time to start a group, recruit scientists, set up experiments, analyse data and publish. There are many factors that need to come together before a group is productive.

For me, finding the right people was essential. Because I wanted to venture into large-scale genomics work on immune cells, I needed to build an interdisciplinary group that would combine expertise in immunology, epigenetics and population genetics. It takes time to find good people.

Once you've recruited, it takes more time to learn how to manage people and navigate a new host institution. Attend leadership courses and seek advice from your senior colleagues if necessary.

Finally, learn to manage expectations. Split tasks into smaller chunks, by setting intermediate goals and working towards achieving them. They will all add up to an overarching goal.

My group is now four years old. The pressure is still present, but I am more in control of it. I now have confidence that my group can design interesting projects, execute experiments, generate high-quality data and analyse them.

Evaluating myself a year from now still worries me, but when the time comes I'll be better prepared. ■

INTERVIEWS BY PAUL SMAGLIK

These interviews have been edited for clarity and length.